

MULTISCALE METABOLIC MODELING OF C4 PLANTS: CONNECTING NONLINEAR GENOME-SCALE MODELS TO LEAF-SCALE METABOLISM IN DEVELOPING MAIZE LEAVES

ELI BOGART AND CHRISTOPHER R. MYERS

ABSTRACT. C4 plants, such as maize, concentrate carbon dioxide in a specialized compartment surrounding the veins of their leaves to improve the efficiency of carbon dioxide assimilation. Nonlinear relationships between carbon dioxide and oxygen levels and reaction rates are key to their physiology but cannot be handled with standard techniques of constraint-based metabolic modeling. We demonstrate that incorporating these relationships as constraints on reaction rates and solving the resulting nonlinear optimization problem yields realistic predictions of the response of C4 systems to environmental and biochemical perturbations. Using a new genome-scale reconstruction of maize metabolism, we build an 18000-reaction, nonlinearly constrained model describing mesophyll and bundle sheath cells in 15 segments of the developing maize leaf, interacting via metabolite exchange, and use RNA-seq and enzyme activity measurements to predict spatial variation in metabolic state by a novel method that optimizes correlation between fluxes and expression data. Though such correlations are known to be weak in general, here the predicted fluxes achieve high correlation with the data, successfully capture the experimentally observed base-to-tip transition between carbon-importing tissue and carbon-exporting tissue, and include a nonzero growth rate, in contrast to prior results from similar methods in other systems. We suggest that developmental gradients may be particularly suited to the inference of metabolic fluxes from expression data.

INTRODUCTION

C4 photosynthesis is an anatomical and biochemical system which improves the efficiency of carbon dioxide assimilation in plant leaves by restricting the carbon-fixing enzyme Rubisco to specialized bundle sheath compartments surrounding the veins, where a high- CO_2 environment is maintained that favors CO_2 over O_2 in their competition for Rubisco active sites, thus suppressing photorespiration [1].

C4 plants are geographically and phylogenetically diverse, and represent the descendants of over 60 independent evolutionary origins of the system [2]. They include major crop plants such as maize, sugarcane and sorghum as well as many weeds and, relative to non-C4 (C3) plants, typically show improved nitrogen and water use efficiencies [3]. The agricultural and ecological significance of the C4 system and its remarkable convergent evolution have made it the object of intense study. The core biochemical pathways are now generally understood [4] but many areas of active research remain, including the genetic regulation of the C4 system [5], the importance of particular components of the system to its function (e.g., [6]), the significance of inter-specific variations in C4 biochemistry [7], details of the process of C4 evolution, [8–12] and the prospect of increasing yields of C3 crops by artificially introducing C4 functionality to those species [13, 14].

Computational and mathematical modeling is a proven approach to gaining insight into C4 photosynthesis and will play an important role in addressing these questions. High-level nonlinear models of photosynthetic physiology [15] relating enzyme activities, light and atmospheric CO₂ levels, and the rates of CO₂ assimilation by leaves have been widely applied to infer biochemical properties from macroscopic experiments and explore the responses of C4 plants under varying conditions. More recently, detailed kinetic models have been used to explore the optimal allocation of resources to enzymes in an NADP-ME type C4 plant [16] and the relationship between the three decarboxylation types [17].

Large-scale constraint-based metabolic models offer particular advantages for the investigation of connections between the C4 system and a plant's metabolism more broadly (for example, partitioning of nonphotosynthetic functions between mesophyll and bundle sheath, or the evolutionary recruitment of nonphotosynthetic reactions into the C4 cycle) and for interpreting high-throughput experimental data from C4 systems. Photosynthesis is difficult to describe, however, using the standard approach of flux balance analysis (FBA), which predicts reaction rates v_1, v_2, \dots, v_N in a metabolic network by optimizing a biologically relevant function of the rates subject to the requirement that the system reach an internal steady state,

$$(1) \quad \begin{aligned} & \max_{(v_1, v_2, \dots, v_N) \in \mathbb{R}^N} f(\mathbf{v}) \\ & \text{s.t.} \quad S \cdot \mathbf{v} = \mathbf{0}, \end{aligned}$$

where the stoichiometry matrix S is determined by the network structure [18]. The relationship between the rate v_c of carbon fixation by Rubisco and the rate v_o of the Rubisco oxygenase reaction depends nonlinearly on the ratio of the local oxygen and carbon dioxide concentrations (here expressed as equivalent partial pressures),

$$(2) \quad \frac{v_o}{v_c} = \frac{1}{S_R} \frac{P_{O_2}}{P_{CO_2}}$$

where S_R is the specificity of Rubisco for CO₂ over O₂. In the C4 case, the CO₂ level in the bundle sheath compartment is itself a function of the rates of the reactions of the C4 carbon concentration system and the rate of diffusion of CO₂ back to the mesophyll.

With the addition of (2), the problem (1) becomes nonlinear and cannot be solved with typical FBA tools; instead (as the problem is also nonconvex [19]), a general-purpose nonlinear programming algorithm is required to numerically solve it.

Prior constraint-based models of plant metabolism have typically ignored the constraint (2) or assumed the oxygen and carbon dioxide levels P_{O_2} and P_{CO_2} are known and fixed v_o/v_c accordingly [20, 21]. While this approach is suitable for mature C4 leaves under many conditions, where v_o/v_c is approximately zero, it may break down in some of the most important targets for simulation: developing tissue, mutants, and C3-C4 intermediate species, where P_{CO_2} in the bundle sheath compartment is not necessarily high.

In other recent work, a high-level physiological model was used to determine v_o , v_c , and other key reaction rates given a few parameters, which were then fixed in order to solve eq. (1) [11]. This method yields realistic solutions, but its application is limited by the lack of a way to set the necessary phenomenological parameters

(e.g., the maximum rate of PEP regeneration in the C4 cycle) based on lower-level, per-gene data (e.g., from transcriptomics or experiments on single-gene mutants).

Here, we treat the problem in a more general way by incorporating the nonlinear constraint (2) directly into the optimization problem (1) and solving the resulting nonlinear program numerically with the IPOPT package [22], using a new computational interface that we have developed, which allows rapid, interactive development of nonlinearly-constrained FBA problems from metabolic models specified in SBML format [23].

Using a new genome-scale reconstruction of the metabolic network of *Zea mays*, developed with particular attention to photosynthesis and related processes, we confirm that this approach can reproduce the nonlinear responses of well-validated, high-level physiological models of C4 photosynthesis [15], while also providing detailed predictions of fluxes throughout the network.

Finally, we combine the results of enzyme assay measurements and multiple RNA-seq experiments and apply a new method to infer the metabolic state at points along a developing maize leaf (Fig. 1a) using a model of mesophyll and bundle sheath tissue in fifteen segments of the leaf, interacting through vascular transport of sucrose, glycine, and glutathione. We compare our results to radiolabeling experiments.

RESULTS

Metabolic reconstruction of *Zea mays*. A novel genome-scale metabolic model was generated from version 4.0 of the CornCyc metabolic pathway database [25] and is presented in two forms. The comprehensive reconstruction involves 2720 reactions among 2725 chemical species, and incorporates CornCyc predictions for the function of 5204 maize genes, with 2064 reactions associated with at least one gene. A high-confidence subset of the model, excluding many reactions not associated with manually curated pathways or lacking computationally predicted gene assignments as well as all reactions which could not achieve nonzero flux in FBA calculations, involves 635 reactions among 603 species, with 469 reactions associated with a total of 2140 genes.

Both the comprehensive and high-confidence models can simulate the production of all major maize biomass constituents (including amino acids, nucleic acids, fatty acids and lipids, cellulose and hemicellulose, starch, other carbohydrates, and lignins, as well as chlorophyll) under either heterotrophic or photoautotrophic conditions and include chloroplast, mitochondrion, and peroxisome compartments, with key reactions of photosynthesis (including a detailed representation of the light reactions), photorespiration, the NADP-ME C4 cycle, and mitochondrial respiration localized appropriately. Gene associations for reactions present in more than one subcellular compartment have been refined based on the results of subcellular proteomics experiments and computational predictions (as collected by the Plant Proteomics Database, [26]) to assign genes to reactions in appropriate compartments.

A model for interacting mesophyll and bundle sheath tissue in the leaf was created by combining two copies of the high-confidence model, with transport reactions to represent oxygen and CO₂ diffusion and metabolite transport through the plasmodesmata, and restricting exchange reactions appropriately (nutrient uptake from

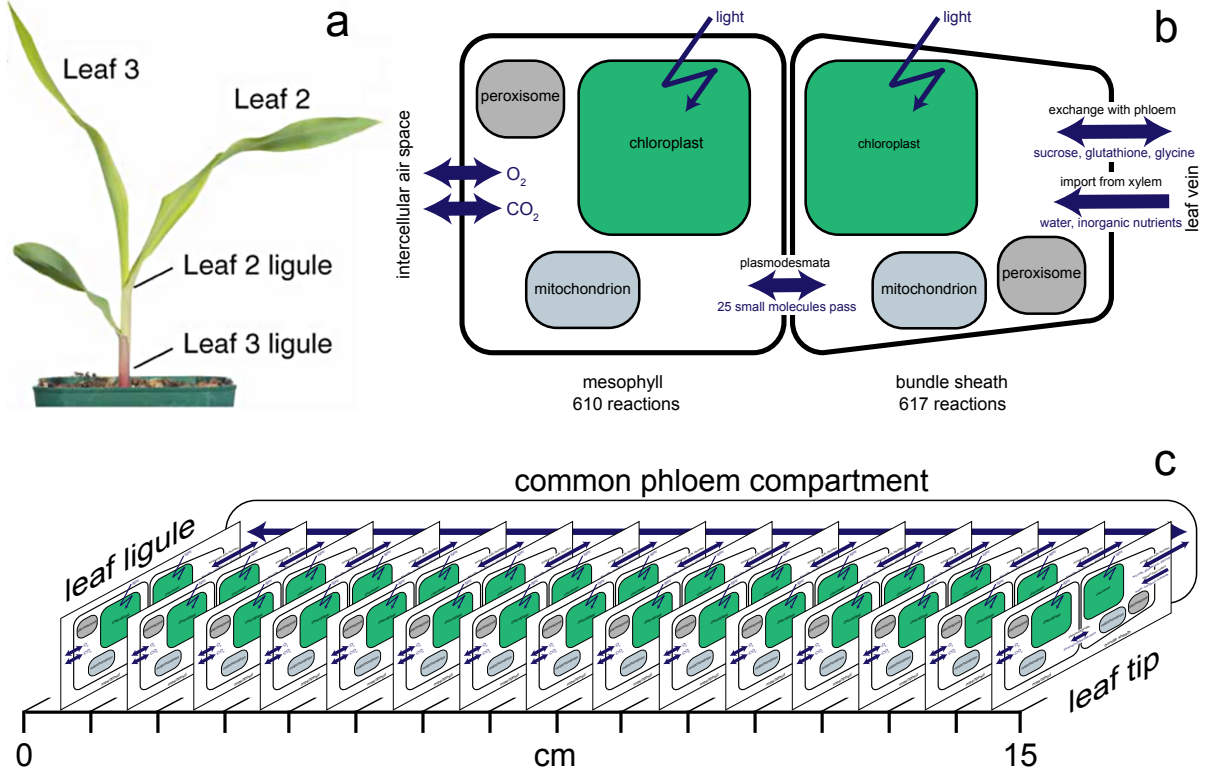


FIGURE 1. **Maize plant and models.** (a) Nine-day-old maize plant (image from [24]). (b) Organization of the two-cell-type metabolic model, showing compartmentalization and exchanges across mesophyll and bundle sheath cell boundaries. (c) Combined 121-compartment model for leaf 3 at the developmental stage shown in (a). Fifteen identical copies of the model shown in (b) represent 1-cm segments from base to tip.

the vascular system to the bundle sheath, and gas exchange with the intercellular airspace to the mesophyll). A schematic of the two-cell model is shown in Fig. 1b.

Both single-cell versions of the model and the two-cell model, designated iEB5204, iEB2140, and iEB2140x2 respectively (based on the primary author's initials and number of genes included, according to the established naming convention [27]), are available in SBML format (S11 Model-S13 Model.)

Nonlinear flux-balance analysis. To solve nonlinear optimization problems incorporating the constraints discussed above, we developed a Python package which – given a model in SBML format, arbitrary nonlinear constraints, a (potentially nonlinear) objective function, and all needed parameter values – infers the conventional FBA constraints of eq. (1) from the structure of the network, automatically generates Python code to evaluate the objective function, all constraint functions, and their first and second derivatives, and calls IPOPT through the pyipopt interface [28]. Source code for the package is available in S14 Protocol and online

(<http://github.com/ebogart/fluxtools>). The software has been used to successfully solve nonlinear FBA problems with over 84000 variables and 62000 constraints.

Figure 2 demonstrates that, as expected, optimizing the rate of CO₂ assimilation in the two-cell-type model with nonlinear kinetic constraints [eqs. (5), (6), (7)] produces predictions consistent with the results of the physiological model of [15]. Note that the effective value of one macroscopic physiological parameter may be governed by many microscopic parameters in the genome-scale model. In the figure, the effective maximum PEP regeneration rate V_{pr} is controlled by the maximum rate of three decarboxylase reactions in the bundle sheath compartment, but with an appropriate choice of parameter values any of at least 10 reactions of the C4 system could become the rate-limiting step in PEP regeneration, and in the calculations below, expression levels for any of the 42 genes associated with these reactions (Supporting Table 1) could influence the net PEP regeneration rate.

Flux predictions in the developing leaf based on multiple data channels.

Maize leaves display a developmental gradient along the base-to-tip direction, with young cells in the immature base and fully differentiated cells at the tip [24, 29]. To explore variations in metabolic state along this axis, we combined the RNA-seq datasets of Wang et al. [30] and Tausta et al. [31] to estimate expression levels (as FPKM) for 39634 genes in the mesophyll and bundle sheath cells at 15 points, representing 1 cm segments of the third leaf of a 9-day-old maize plant, which includes a full gradient of developmental stages. The combined dataset provides expression information for 920 reactions in the two-cell model (460 each in mesophyll and bundle sheath cells).

A whole-leaf metabolic model, iEB2140x2x15, was created from fifteen copies of the two-cell model, each representing a 1-cm segment, interacting through the exchange of sucrose, glycine, and glutathione through a common compartment representing the phloem. The resulting 121-compartment model, Fig. 1c, involves 18780 reactions among 16575 metabolites.

Subject to the requirements that reaction rates in each of the 15 segments obey both the FBA steady-state constraints (eq. 1) and the nonlinear constraints governing Rubisco kinetics (eqs. 5, 7, and 6, presented in detail below) we determined the set of rates v_{ij} for each reaction i at each segment j which were most consistent with the base-to-tip variation in the gene expression data, by optimizing the objective function

$$(3) \quad F(v) = \sum_{i=0}^{N_r} \sum_{j=1}^{15} \frac{(e^{s_i} |v_{ij}| - d_{ij})^2}{\delta_{ij}^2} + \alpha \sum_{i=0}^{N_r} s_i^2$$

where $N_r = 920$ is the number of reactions associated with at least one gene present in the expression data, d_{ij} and δ_{ij} are the expression data and associated experimental uncertainty for reaction i at leaf segment j , and s_i is an optimizable scale factor associated with reaction i .

Effectively, this calculation – similar to the method of Lee et al. [32] or FALCON [33] – performs a constrained least-squares fit of the fluxes to the expression data. Allowing the scale factors s_i to vary emphasizes agreement between fluxes and data in their trend along the developmental gradient, rather than in their absolute value: if the data associated with reaction R_i has average value 100 FPKM, a solution in which R_i has mean flux 10 $\mu\text{mol m}^{-2} \text{s}^{-1}$ but correlates well with the data

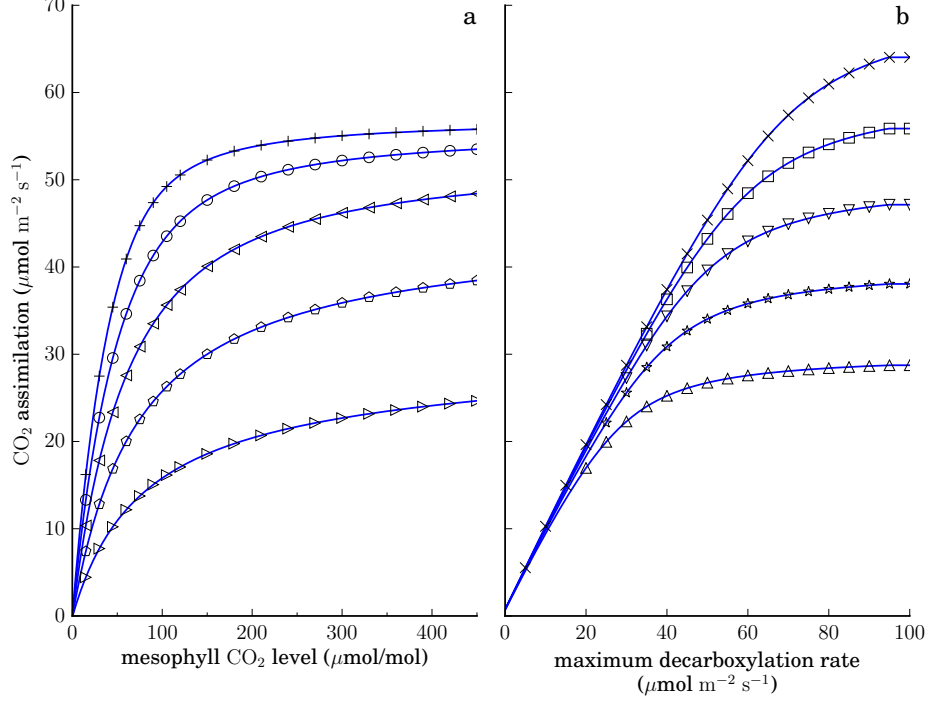


FIGURE 2. **CO₂ assimilation rates (A) predicted by the C4 photosynthesis model of [15], solid lines, and the present nonlinear genome-scale model (markers) maximizing CO₂ assimilation with equivalent parameters.** Left, A vs mesophyll CO₂ levels with varying PEPC levels (top to bottom, $V_{p,\max} = 110, 90, 70, 50,$ and $30 \mu\text{mol m}^{-2} \text{s}^{-1}$). Right, A vs total maximum activity of all bundle sheath decarboxylase enzymes (equivalent to the maximum PEP regeneration rate V_{pr}) at varying Rubisco levels (top to bottom, $V_{c,\max} = 70, 60, 50, 40,$ and $30 \mu\text{mol m}^{-2} \text{s}^{-1}$). Other parameters as in Table 4.1 of [15], except with nonphotorespiratory respiration rates $r_d = r_m = 0$.

can achieve (with appropriate choice of scale factor) a lower cost than a solution in which R_i has mean flux $100 \mu\text{mol m}^{-2} \text{s}^{-1}$ but is anticorrelated. The penalty term $\alpha \sum s_i^2$ favors solutions in which, generally, reactions with larger associated expression data carry higher fluxes. The parameter α controlling the tradeoff between these criteria was set arbitrarily to 1.0 in the work presented here. We require $s_a = s_b$ if reactions a and b are mesophyll and bundle sheath instances of the same reaction.

To constrain the overall scale of the fluxes and further improve accuracy, we incorporated enzyme activity assay data from [30] for seventeen enzymes (including Rubisco and PEPC) along the 15 leaf segments as additional constraints on the optimization problem, requiring for each enzyme k and segment j

$$(4) \quad E_{jk} \geq |v_{k1}| + \dots + |v_{kn}|$$

where E_{jk} is the measured maximal activity of the enzyme at that segment and the sum on the right hand side includes all the reactions which represent enzyme k in the mesophyll, bundle sheath, and subcompartments of those cells if applicable.

Solving the optimization problem yielded predictions for reaction rates and other variables (S16 Table). Upper and lower bounds on selected variables (S17 Table) were determined through flux variability analysis (FVA) [34], allowing the objective function to increase by 0.1% from its optimal value.

Predicted source-sink transition. As shown in Fig. 3, in the outer, more photosynthetically developed, portion of the leaf, our optimal fit predicts net CO_2 uptake, with most of the assimilated carbon incorporated into sucrose and exported to the phloem. Near the base of the leaf, sucrose is predicted to be imported from the phloem and used to drive a high rate of biomass production, with some concomitant net release of CO_2 to the atmosphere by respiration.

This transition between a carbon-exporting source region and a carbon-importing sink region is well known, and the predicted transition point between the two, approximately 6 cm above the base of the leaf, can be compared to the ^{14}C -labeling results of Li et al. [24] in the same experimental conditions. Fig. 3b shows the location of labeled carbon in leaf 3 after feeding labeled CO_2 to leaf 2 (center image) or leaf 3 (bottom image). Li et al. [24] identified the sink region as the lowest 4 cm of the leaf; the transition is not perfectly sharp and quantitative comparison of exchange fluxes is not possible, but the nonlinear FBA results appear to slightly overestimate the size of the sink region.

Agreement might be improved under a different assumption about net sucrose import or export by leaf 3 (here, we have assumed that the import visible in the center image is exactly balanced by the export suggested by the high density of labeled carbon at the absolute base in the lower image.)

The net rate of CO_2 assimilation predicted in the outer, most mature leaf segments, $8\text{--}11 \mu\text{mol m}^{-2} \text{s}^{-1}$, is lower than that typically measured in more mature maize plants (e.g., rates of $20\text{--}30 \mu\text{mol m}^{-2} \text{s}^{-1}$ in 22-day-old wild-type plants under comparable conditions [6]), but photosynthetic capacity may still be increasing even in these segments.

In addition to sucrose, glycine and glutathione are predicted to be exported from the source region through the phloem and reimported by the sink region, consistent with our expectations that nitrogen and sulfur reduction will occur preferentially in the photosynthesizing region (Supporting Figure 1). Note that this behavior emerges from the data even though there is no explicit requirement in the model that net phloem transport occur in a basipetal direction.

Predicted C4 system function. Figure 4 shows predicted rates of key reactions of the C4 system and CO_2 and O_2 levels in the bundle sheath. As expected, the model predicts that a C4 cycle will operate in the source region of the leaf, elevating the CO_2 level in the bundle sheath. The CO_2 level is also elevated in the source region; this is an immediate consequence of respiration in the bundle sheath and eq. (7). It may be overestimated here because we have assumed a constant value for the bundle sheath CO_2 conductivity (as measured by Bellasio et al. [35]); in fact, gene expression associated with synthesis of the diffusion-resistant suberin layer between bundle sheath and mesophyll peaks at 4 cm above the leaf base [30], so g_s is presumably higher below that point.

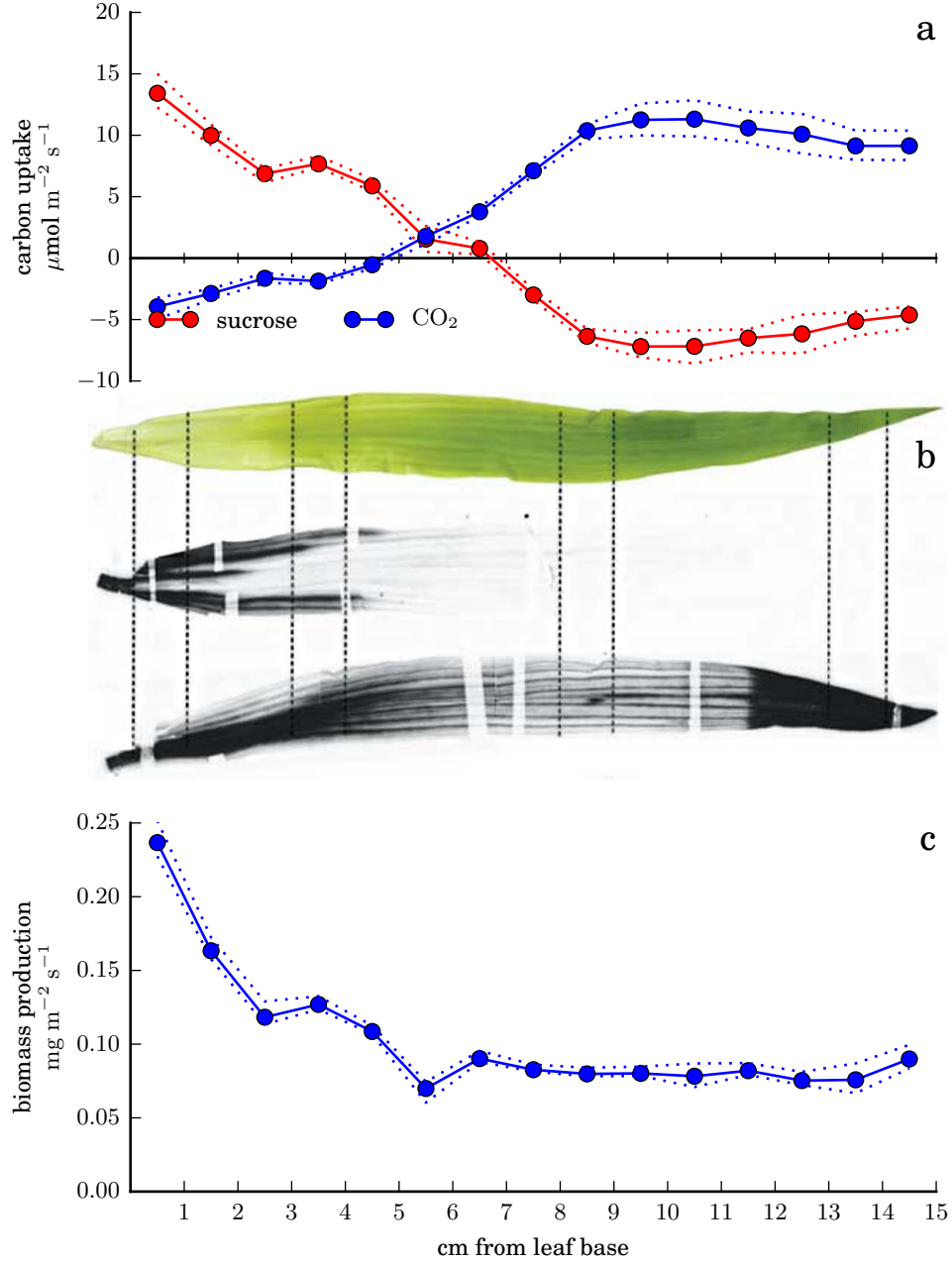


FIGURE 3. **Source-sink transition along the leaf as predicted by optimizing the agreement between fluxes in the nonlinear model and RNA-seq data.** (a) Predicted rates of exchange of carbon with the atmosphere and phloem along the leaf. (b) Experimental observation of the source-sink transition, reproduced from [24]. Upper image, photograph of leaf 3; middle image, autoradiograph of leaf 3 after feeding $^{14}\text{CO}_2$ to leaf 2; lower image, autoradiograph of leaf 3 after feeding $^{14}\text{CO}_2$ to the tip of leaf 3. (c) Total biomass production in the best-fitting solution. In panels a and c, dotted lines indicate minimum and maximum predicted rates consistent with an objective function value no more than 0.1% worse than the optimum.

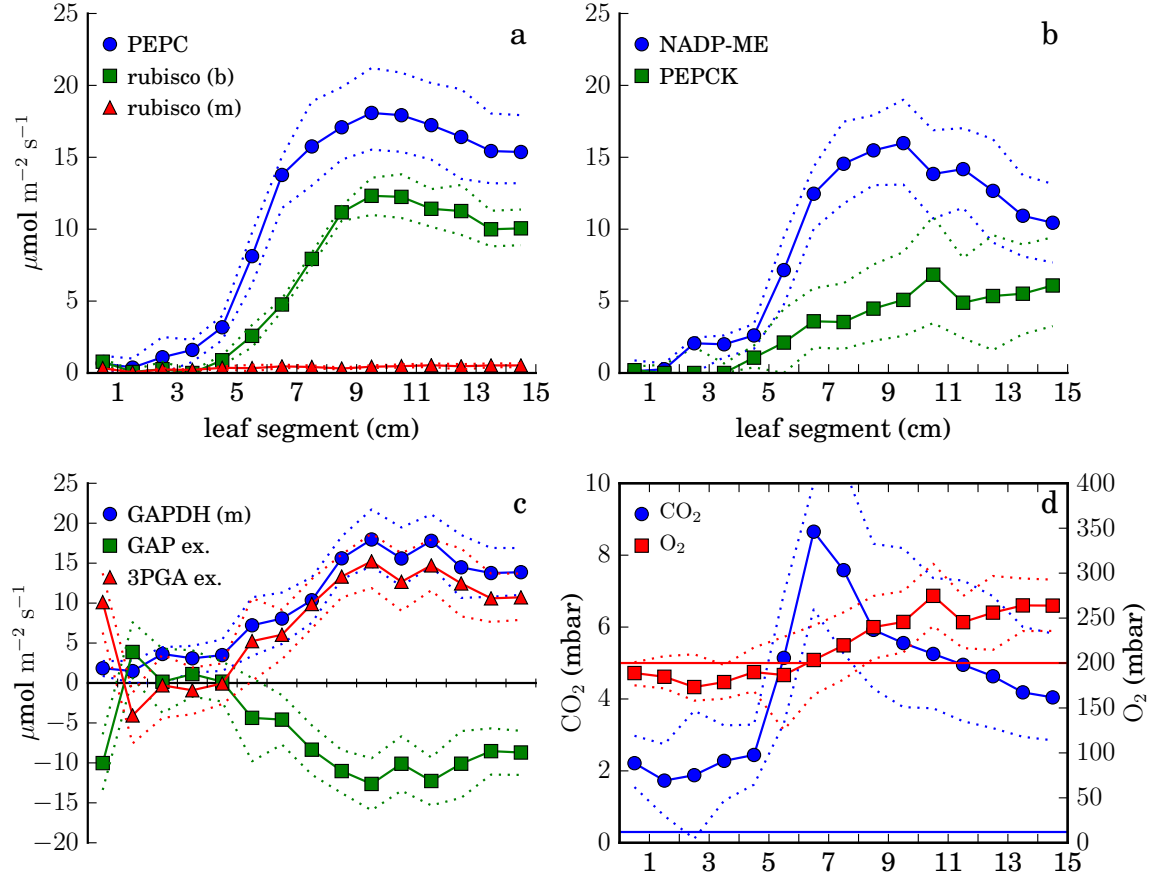


FIGURE 4. **Operation of the C4 system in the best-fitting solution.** (a) Rates of carboxylation by PEPC in the mesophyll and Rubisco in the mesophyll and bundle sheath. (b) Rates of CO_2 release by PEP carboxykinase and chloroplastic NADP-malic enzyme in the bundle sheath. (c) Transport of 3-phosphoglycerate and glyceraldehyde 3-phosphate from bundle sheath to mesophyll (or the reverse, where negative) and glyceraldehyde 3-phosphate dehydrogenation rate in the mesophyll chloroplast, showing the involvement of the mesophyll in the reductive steps of the Calvin cycle throughout the source region. (d) Oxygen and carbon dioxide levels in the bundle sheath. Straight lines show mesophyll levels. Throughout, dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.

In the Calvin cycle, most reactions are predicted to be bundle-sheath specific, but the reductive phase is active in both cells, with approximately half the 3-phosphoglycerate produced in the bundle sheath transported to the mesophyll and returned as dihydroxyacetone phosphate (Fig. 4c); this is a known aspect of NADP-ME C4 metabolism connected to reduced photosystem II activity in the bundle sheath cells [36], which is also predicted here (Supporting Figure 2). Consistent with conclusions drawn independently from the transcriptomic data, as well as proteomic data from the same system [24, 30, 37], the model does not predict a C3-like metabolic state as a developmental intermediate stage. As expected in maize [38], a significant role for phosphoenolpyruvate carboxykinase (PEPCK) as a decarboxylating enzyme operating in the bundle sheath in parallel with NADP-ME is predicted (Fig. 4b).

While the predictions are generally consistent with the standard view of the C4 system in maize, there are minor discrepancies. In the mesophyll, our calculations predict that malate production occurs in the mitochondrion, rather than the chloroplast. In both mesophyll and bundle sheath, phosphoenolpyruvate is formed by pyruvate-orthophosphate dikinase (PPDK) in the chloroplast at a higher rate than necessary to sustain the C4 cycle; the excess is converted again to pyruvate by pyruvate kinase in the cytoplasm, with the resulting ATP consumed by the model’s generic ATPase reaction. Finally, in the bundle sheath, a modest rate of PEPCK activity is predicted, recapturing CO_2 only to have it released again by the decarboxylases (S3 Figure). Further refinement of the associations of genes to reactions in the model might resolve some of these discrepancies.

Global agreement between fluxes and data. Figure 5 summarizes overall properties of the predicted fluxes. It is not clear why agreement between data and predicted fluxes is poorer at the base, as shown in Fig. 5a. As discussed below, the cell-type-specific RNA-seq data from Tausta et al. [31] does not extend below the fourth segment from the base of the leaf; at the base we have assumed expression levels for all genes are equal in mesophyll and bundle sheath. Though proteomics experiments on the same system [37] generally found limited cell-type specificity at the leaf base, this assumption is likely an oversimplification, and could limit the ability of the algorithm to find a flux prediction consistent with the data there.

For most reactions, the correlation between the base-to-tip expression pattern and the base-to-tip trend in predicted flux is high. The cumulative histogram in Fig. 5b shows that the Pearson correlation $r > 0.92$ for more than half of the reactions in the model with associated expression data.

Differences in expression levels between different reactions, however, correlate only weakly with the differences in fluxes between those reactions, as shown for segment 15 in Fig. 5c (blue circles). After rescaling fluxes by the optimal per-reaction scale factors, a clear relationship emerges (Fig. 5c, red circles), confirming that the scale factors are functioning as intended. Of course we should not expect a perfect correlation between data on transcript levels and predicted fluxes through associated reactions. The limited correlation between fluxes and expression data across different reactions presumably follows, in part, from the imperfect correlation between expression data and protein abundance across different genes, as illustrated in Fig. 5d with data from the same experimental system [39], as well as from the different catalytic capabilities of different enzymes, posttranslational regulation, differences in substrate availability, etc.

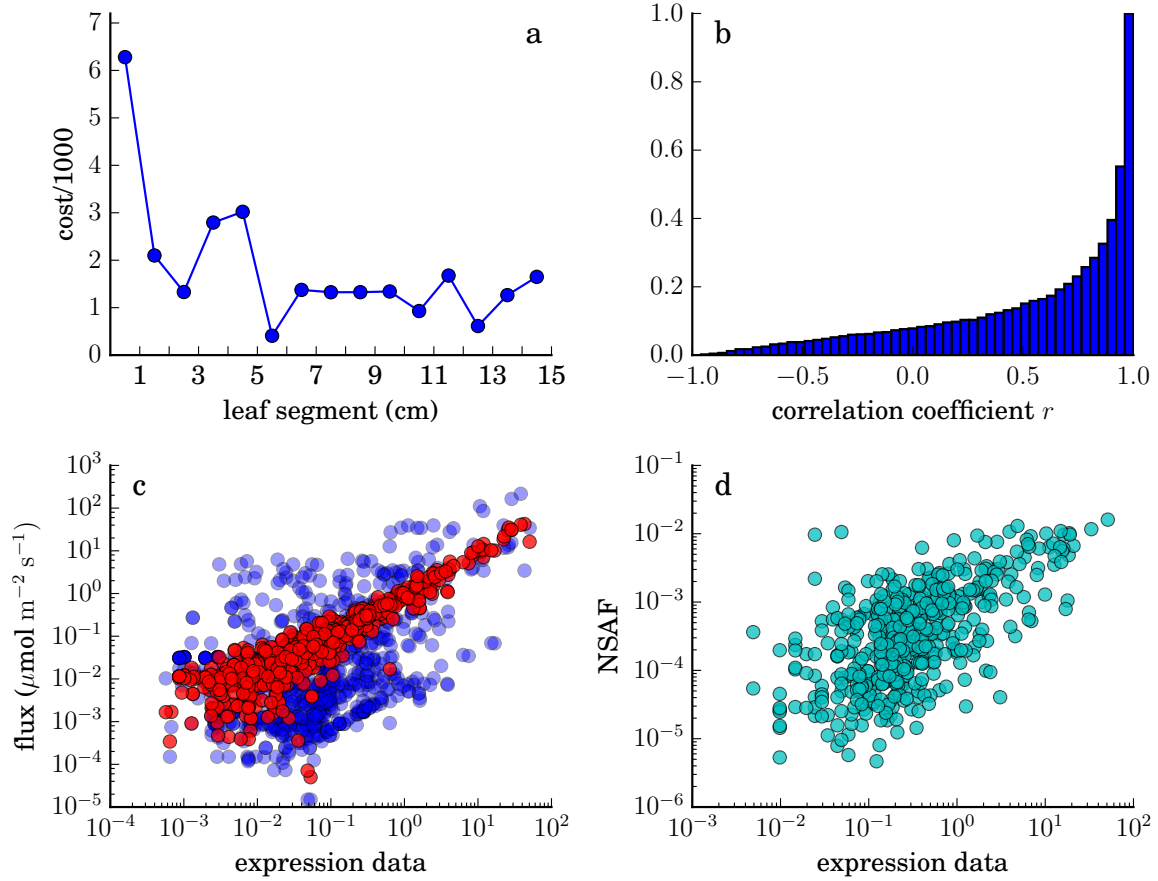


FIGURE 5. **Agreement between RNA-seq data and predicted fluxes.** (a) Contribution of each segment to the objective function (eq. (3), excluding costs associated with scale factors). (b) Cumulative histogram of Pearson correlations between data and predicted fluxes for all reactions. (c) Predicted fluxes versus expression data at the tip of the leaf (blue, raw fluxes; red, after rescaling each flux v_i by the optimal factor e^{s_i} of eq. (3)). Some outliers with very low predicted flux are not shown. (d) Relationship between RNA-seq and proteomics measurements for 506 proteins in the 14th segment from the base, redrawn from the data of [39]. NSAF, normalized spectral abundance factor.

Reconciling expression data and network structure. Figure 6 illustrates the operation of the fitting algorithm in detail, using two regions of the metabolic network with simple structure as examples.

In Fig. 6a, expression data for eight reactions of the pathway leading to chlorophyllide a are shown. Expression levels for the different reactions at any point on the leaf may span an order of magnitude or more, but the FBA steady-state assumption requires the rates of all reactions in this unbranched¹ pathway to be equal at each point. Applying the optimal rescaling determined for each reaction’s expression data, shown in panel b, allows the flux prediction for the pathway (solid dots) to achieve reasonable agreement with the data. (Note that data for reaction 4 cannot be further scaled down because of the lower limit $\exp(-5)$ on its scale factor $\exp(s_4)$, imposed for technical reasons.)

Figure 6c shows data for a three-reaction branch point in aromatic amino acid synthesis. To balance production and consumption of arogonate, the prephenate transaminase flux must equal the sum of the fluxes through arogonate dehydrogenase (to tyrosine) and arogonate dehydratase (to phenylalanine) but expression is consistently lower for the transaminase than the other enzymes. After rescaling (Fig. 6d), the data agree well with the stoichiometrically consistent flux predictions (solid dots). The predicted ratio of dehydrogenase to dehydratase flux reflects data for downstream reactions.

Comparison to other methods for integrating RNA-seq data. Supporting Figure 4 shows predictions that result when the scale factors s_i of eq. (3) are fixed to zero. The source-sink transition is apparent but the C4 cycle operates at lower levels, the example pathways of Fig. 6 (and a number of others) show little or no activity, and predicted fluxes along the leaf are not as tightly correlated with their associated expression data.

Supporting Figure 5 shows the metabolic state predicted by applying the expression data for each reaction as an upper bound on the absolute value of the reaction rate as in the E-Flux method [40] to the fifteen-segment model with the same RNA-seq data. The C4 system is predicted to operate, but no source-sink transition is apparent, and typical data-predicted flux correlations are poor. Imposing a realistic biomass composition restores the source-sink transition and somewhat improves correlation between data and fluxes (Supporting Figure 6). Fluxes predicted by E-Flux are generally smaller than those predicted by the least-squares method, with or without per-reaction scale factors.

Supporting Figure 9 compares the fluxes predicted at the tip by optimizing agreement with the data through the non-biological objective function (eq. 3), fluxes predicted at the tip with an explicit biological objective function (maximizing CO_2 assimilation) constrained by the experimental data in the E-Flux method, and fluxes predicted in an FBA calculation which ignores the data entirely (minimizing total flux while achieving the same CO_2 assimilation rate as predicted at the tip by the least-squares method.) Both data-integration methods lead to predictions very different from the unconstrained FBA calculation.

¹The branch leading to heme production is not included in the reconstruction.

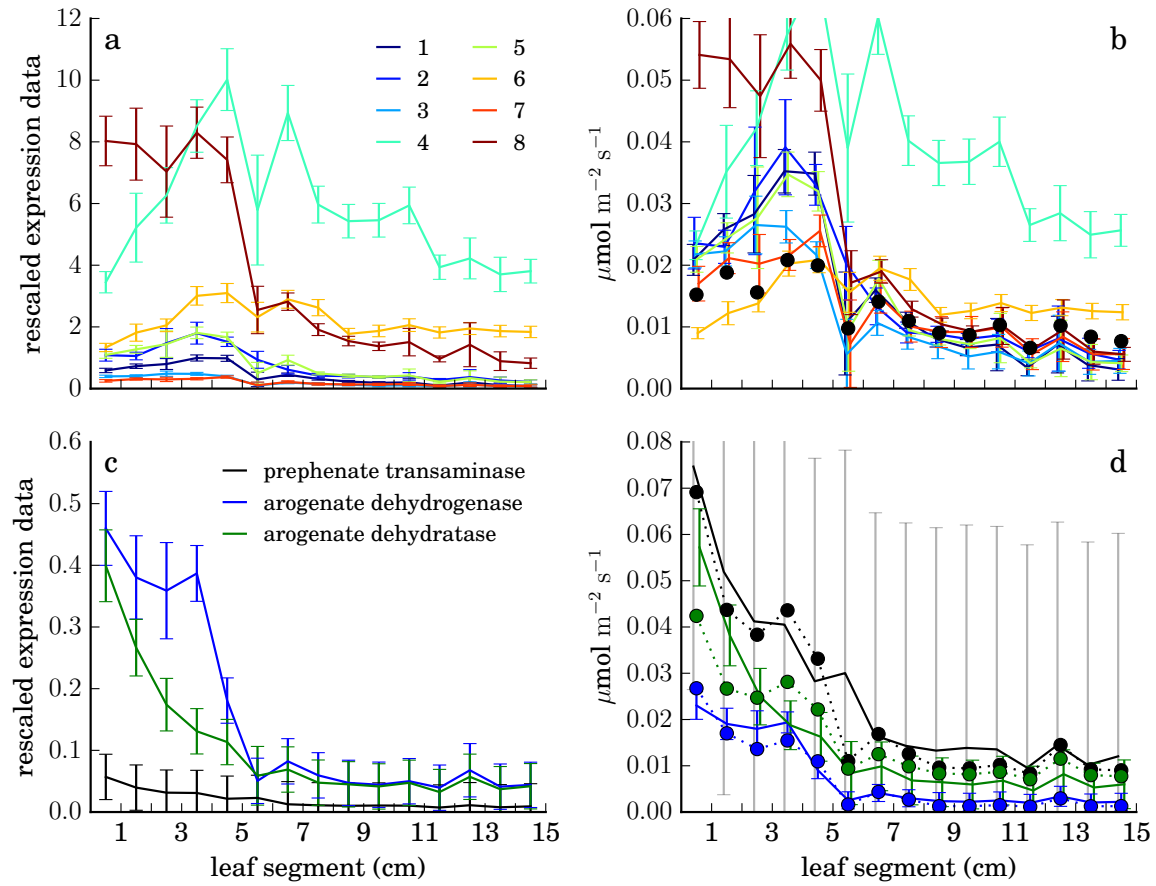


FIGURE 6. Comparison of RNA-seq data to predicted fluxes for a linear pathway and around a metabolic branch point. Upper panels, chlorophyllide a synthesis in the mesophyll; lower panels, production of arogenate in the bundle sheath by prephenate transaminase and its consumption by arogenate dehydrogenase and arogenate dehydratase. Left, aggregate RNA-seq data and experimental standard deviations for each reaction rescaled by a uniform factor (see text). Right, same data and errors further rescaled by reaction-specific optimal factors (e^{-s_i} , in the variables of eq. 3) to best match data with predicted fluxes (solid circles). Fluxes are equal for all reactions of the linear pathway (1, uroporphyrinogen decarboxylase, 2, coproporphyrinogen oxidase, 3, protoporphyrinogen oxidase, 4, magnesium chelatase, 5, magnesium protoporphyrin IX methyltransferase, 6, magnesium protoporphyrin IX monomethyl ester cyclase, 7, divinyl chlorophyllide a 8-vinyl-reductase, 8, protochlorophyllide reductase.) Error bars represent standard deviations of expression measurements across multiple replicates.

DISCUSSION

Reconstruction. Our model is the fourth published genome-scale metabolic reconstruction of the major crop plant *Zea mays*, and the first such reconstruction developed solely from maize data sources, rather than as a direct or indirect adaptation of the *Arabidopsis thaliana* model AraGEM [20].

Direct reaction-to-reaction comparison of iEB5204 with C4GEM [41], iRS1563 [21], and its successor model [42] is difficult because those models use a naming scheme for compounds and reactions ultimately based on KEGG [43, 44] while this model, like its parent database, uses the nomenclature of MetaCyc and the BioCyc database collection. The models are broadly similar in size and biological scope. As published, C4GEM included 1588 reactions associated with 11623 maize genes; iRS1563, 1985 reactions associated with 1563 genes; the model of Simons et al. [42], 3892 unique reactions and 5824 genes; and iEB5204, 2720 reactions with 5204 genes. All models can simulate the production of similar sets of basic biomass constituents (including amino acids, carbohydrates, nucleic acids, lipids and fatty acids, and cell wall components) under photosynthetic and non-photosynthetic conditions and include key reactions of the C4 cycle. The model of Simons et al. [42] also offers extensive coverage of secondary metabolism.

However, the present model has several advantages which make it particularly suitable for integration with transcriptomics data:

Gene associations: The gene associations included in iEB5204 are those presented in CornCyc [25], which are generated by the PMN Ensemble Enzyme Prediction Pipeline (E2P2) [45], a homology-based protein sequence annotation algorithm trained on a reference dataset of experimentally validated enzyme sequences. The E2P2 approach is more comprehensive and scalable than the development procedures of the previous maize reconstructions (which involve, for example, obtaining gene associations by transferring annotations from Arabidopsis genes to their best maize BLAST hits and manually selecting annotations for remaining maize genes from among BLAST hits in other species.) The entire set of gene associations in the FBA model may be readily updated based on improvements in the E2P2 prediction algorithm.

High-confidence submodel: In developing the fitting algorithm we found that, to obtain plausible metabolic state predictions, a conservative reconstruction was preferable to a comprehensive one. For example, early tests with the comprehensive version of the model suggested that the fitting algorithm often found low-cost solutions involving high fluxes through reactions which, on investigation, we determined were unlikely to be active in maize. Because of the model’s connection to the CornCyc database, it was straightforward to create a reduced, high-confidence version of the model by preferentially excluding reactions not included in any manually curated plant metabolic pathway, even if candidate associated genes had been identified computationally, leading to more realistic results.

Reproducibility: In an effort to improve the reusability of the model and encourage its application to other data sets, we have provided the full source code (S14 Protocol and S15 Protocol) for all calculations presented here, as has been recommended (see, e.g., [46]).

Previous reconstructions do offer two features absent from this model: gene associations for intracellular transport reactions, and gene associations which take into account the structure of protein complexes. Both should be considered in future work.

In agreement with [47], we found that building the model starting from a metabolic pathway database was considerably more straightforward than the standard process of *de novo* reconstruction [48]. Reasonable effort was still required to bring the model to a functional state by identifying reactions or pathways present in the CornCyc database which could not be handled automatically by the Pathway Tools export facility (for example, because they involved polymerization, or could not be checked automatically for conservation violations) and determining how to represent them appropriately in the FBA model.

The model construction process here could readily be adapted to generate metabolic models describing any of the more than 30 crop and model plant species for which Pathway Tools-based metabolic pathway databases [49] have been developed by the Plant Metabolic Network [50], Sol Genomics Network [51], Gramene [52], and others (e.g., [53–55]) allowing the present data-fitting method to be applied to RNA-seq data from those organisms. The level of model development effort required and quality of fit results will vary depending on the extent of curation of the pathway database and quality of the gene function annotations.

Nonlinear optimization. In contrast to the linear and convex optimization methods employed in nearly all prior constraint-based modeling work, general constrained nonlinear optimization algorithms typically require more effort from the user (who might be required to supply functions which evaluate the first and second derivatives of all constraints with respect to all variables in the problem). They are slower, are more sensitive to choices of starting point and problem formulation, are not guaranteed to converge to an optimal point even if one exists, and, when they do converge to an optimum, cannot guarantee that it is globally optimal.

The software package we present allows the rapid and effective development of metabolic models with nonlinear constraints despite these complications. All necessary derivatives of constraint functions are taken analytically, and Python code to evaluate them is automatically generated. A model in SBML format may be imported, nonlinear constraints added and removed, and the problem repeatedly solved to test various design choices, solver options, and initial points, all within an interactive session, with a minimum of initial investment of effort in programming.

In the present case, agreement between nonlinear FBA calculations maximizing growth and the predictions of classical physiological models confirmed that the true, globally optimal CO₂ assimilation rate was found successfully. For the data-fitting calculations, where the true optimal cost is not known, we cannot exclude the possibility that there exist other optimal solutions, qualitatively distinct from the flux distributions and quasi-optimal regions presented above, with equivalent or lower costs. In practice, we encountered occasional cases in which reaction or pathway fluxes were initially predicted to be zero even when associated with nonzero data, despite the existence of a superior alternative solution with nonzero predicted fluxes. A step to detect and correct these situations was incorporated into the fitting algorithm.

Many future applications for the software are possible. Our approach to Rubisco kinetics may easily be extended to other models of C4 metabolism or, more generally, to any FBA calculation in a photosynthetic organism where the CO_2 level at the Rubisco active site, and thus the Rubisco oxygenation/carboxylation ratio, is not known *a priori*. A recent genome-scale metabolic reconstruction of the model alga *Chlamydomonas reinhardtii*, for example, was identified by the authors as being deficient in describing algal metabolism under low CO_2 conditions due to the fact that the Rubisco carboxylase and oxygenase fluxes were treated as independent and not competitive, as we have done here [56].

Ensuring that rates of Rubisco oxygenation, Rubisco carboxylation, and PEPC carboxylation are consistent with our knowledge of their kinetics is a special case of the more general problem of integrating kinetic and constraint-based modeling, to which diverse approaches have been proposed (e.g., [57–62]).

To our knowledge, no prior work has simply imposed kinetic laws as additional, nonlinear constraints in the ordinary FBA optimization problem. Our results demonstrate the potential of this approach in systems where the kinetics of a few well-understood reactions are crucial. It remains to be seen how many kinetic laws may be incorporated in this way at once, and to what extent their introduction usefully constrains the space of possible steady-state flux distributions even when relevant kinetic parameters are not known (but instead are treated as optimizable variables, an approach with connections to ensemble kinetic modeling [63]).

Nonlinear constraints may also be of use in enforcing thermodynamic realizability of flux distributions, and relaxing requirements of linearity or convexity may stimulate the development of novel objective functions – either for data integration purposes, as here, or as alternatives to growth-rate maximization.

Data fitting. The expression of a gene encoding a metabolic enzyme need not correlate with the rate of the reaction that enzyme catalyzes. The relationship between transcription and degradation of mRNA and control of flux is indirect, mediated by protein translation, folding, and degradation, complex formation, posttranslational modification, allosteric regulation, and substrate availability. Indeed, as reviewed by [64], experimentally observed correlations among RNA-seq or microarray data (each itself an imperfect proxy for mRNA abundance or transcription rate), protein abundance, enzyme activity, and fluxes are variable and often weak.

For example, RNA-seq and quantitative proteomic data obtained from maize leaves at the same developmental stage studied here, harvested simultaneously from plants grown together, showed Pearson correlation approximately 0.6 across the entire dataset, but some significantly lower values were found when correlations were restricted to genes of particular functional classes, and measured mRNA/protein ratios for individual genes varied up to 10-fold along the gradient [39]. A subset of this data is shown in Fig. 5d.

The most comprehensive study of the issue in plants so far [65] found so little agreement between RNA-seq and ^{13}C -MFA data from embryos of two *Brassica napus* accessions that the authors concluded the inference of central metabolic fluxes from transcriptomics is, in general, impossible.

In this light, it is not surprising that methods for integrating transcriptomic data with metabolic models to predict reaction rates have met with limited success. Machado and Herrgård [66] reviewed 18 such methods and assessed the performance of seven of them on three test datasets from *E. coli* and *Saccharomyces cerevisiae*

where experimentally measured intracellular and extracellular fluxes were available for comparison. None of the methods consistently outperformed parsimonious FBA simulations which completely ignored transcriptomic data.

In contrast, in the present work the use of transcriptomic data (and a limited number of enzyme activity measurements) allowed the correct prediction of a metabolic transition from the base of the leaf to the tip, which could not have been expected based on FBA calculations alone: without such data, all points along the gradient would be identical, and the biomass-production-maximizing solution would be the same at each. The predicted position of the source-sink transition is not perfectly accurate, and the overall performance of the model cannot be evaluated until the predicted reaction rates are compared to detailed experimental flux measurements. Nonetheless, the results are encouraging. We offer two explanations for this apparent success.

First, the metabolic transition between the heterotrophic sink region at the base and the photoautotrophic source region at the tip is particularly dramatic, involving a large number of reactions which are effectively absent in one region but carry high fluxes in the other [24]; so long as even a slight correlation between transcript levels and fluxes exists, such a reconfiguration should be apparent from expression data.

Second, although the developing maize leaf is biologically more complex than microbial growth experiments, the relationship between expression levels and fluxes may be actually be closer in the leaf. Leaf development is a stereotyped, frequently repeated, relatively slow, one-way process, in which the precise sequence of events is subject to evolutionary optimization. Coordination of transcription with required fluxes will lead to efficient use of resources. In contrast, the test cases of [66] involve microbial responses to varying environmental conditions and under- and over-expression mutations. Environmental responses must be rapid, flexible and reversible – criteria a complex, scripted transcriptional response may not satisfy – while transcriptional responses to novel mutations, by definition, cannot have been evolutionarily optimized. This hypothesis could be tested by evaluating performance of the present method on RNA-seq data from mutant maize plants, or plants subject to environmental challenges.

We note also that methods that did not constrain or optimize the growth rate predicted zero growth rates in almost all the test cases studied by Machado and Herrgård [66]. The present method also does not constrain or optimize the growth rate but consistently does predict nonzero growth as reflected in nonzero biomass production (whether with a flexible biomass composition was used, as above, or a fixed biomass composition, as in Supporting Figure 7 and Supporting Figure 8).

The whole-leaf model. Large-scale metabolic models of interacting cells of multiple types first appeared in 2010, with C4GEM [41] and a model of human neurons interacting with their surrounding astrocytes [67]. Many more complex multicellular FBA models have since appeared, including studies of the metabolism of interacting communities of microbial species in diverse natural environments or artificial co-cultures [68–74] (also [75] at a smaller scale) and of the metabolic capacities of host animals and their symbionts [76] or parasites [77]. In plants, diurnal variation in C3 and CAM plant metabolism has been simulated with a model which represents different phases of the diurnal cycle with different abstract compartments, with transport reactions representing accumulation of metabolites over time [78].

In the most direct antecedent of the present work, Grafahrend-Belau and coauthors developed a multiscale model of barley metabolism [79] which represented leaf, stem, and seed organs as subcompartments of a whole-plant FBA model, with nutrients exchanged through the phloem. Combining the FBA model with a high-level dynamic model of plant metabolism allowed them to predict changes in metabolism over time, including the transition between a biomass-producing sink state and a fructan-remobilizing source state in the stem late in the plant’s life cycle.

The whole-leaf model presented here occupies an intermediate position between prior C4 models, with single mesophyll and bundle sheath cells, and multi-organ whole-plant models such as [79]. It represents the first attempt to model spatial variations in metabolic state within a single organ, allowing the study of developmental transitions in leaf metabolism by incorporating data from more and less differentiated cells at a single point in time, rather than modeling development dynamically.

Other interacting cell models incorporate *a priori* qualitative differences in the metabolic capabilities of their components (e.g., leaf, stem, and seed, or neurons and astrocytes). In contrast in the work presented here, in order to allow the metabolic differences between any two adjacent points to be purely quantitative, the same metabolic network must be used for all points. This simplifies the process of model creation but implies that meaningful predictions of spatial variation depend entirely on the integration of (spatially resolved) experimental data. The ability of the model to capture the experimentally observed shift from sink to source tissue along the developmental gradient based on RNA-seq and enzyme activity measurements shows that this may be done successfully with high-resolution -omics data and careful model construction.

METHODS

Reconstruction process. A local copy of CornCyc 4.0 [25] was obtained from the Plant Metabolic Network and a draft metabolic model was created using the MetaFlux module of Pathway Tools 17.0 [47]. The resulting model, including reaction reversibility information, was converted to SBML format and iteratively revised, as described in detail in Outline, until all desired biomass components could be produced under both heterotrophic and photosynthetic conditions and realistic mitochondrial respiration and photorespiration could operate.

An overall biomass reaction was adapted from iRS1563 [21] with minor modifications to components and stoichiometry, as detailed in Outline. To allow calculations with flexible biomass composition, individual sink reactions were added for most species participating in the biomass reaction, as well as several relevant species (including chlorophyll) not originally included in the iRS1563 biomass equation.

Core metabolic pathways were assigned appropriately to subcellular compartments (e.g., the TCA cycle and mitochondrial electron transport chain to the mitochondrion; the light reactions of photosynthesis, the Calvin cycle, and some reactions of the C4 cycle to the chloroplast; and some reactions of the photorespiratory pathway to the peroxisome) and the intracellular transport reactions necessary for their operation were added.

The model was thoroughly tested for consistency and conservation violations, confirming that no species could be created without net mass input or destroyed

without net mass output (except species representing light, which can be consumed to drive futile cycles.)

The base metabolic model iEB5204 is provided in SBML format as S11 Model. Gene association rules for reactions with associated genes in CornCyc are provided following COBRA conventions [80]. Additional annotations give the record in the CornCyc database associated with each reaction and species, where applicable.

To produce the higher-confidence version of the reconstruction, iEB2140 (S12 Model), reactions in the base model which were not associated with any identified metabolic pathway in CornCyc, and those for which no genes for a catalyzing enzyme had been identified by computational function prediction, were removed from the model if their removal did not prevent photosynthesis, photorespiration, or the production of any biomass component. Then, all reactions which could not achieve nonzero steady-state rates were removed.

Mesophyll-bundle sheath model. A model for leaf tissue (S13 Model) was created by taking two copies of the high-confidence model, representing mesophyll and bundle sheath cells, and adding reactions representing transport through the plasmodesmata which connect the cytoplasmic spaces of adjacent cells. Though in principle most small molecules can cross the plasmodesmata by diffusion [81], unrealistic concentration gradients may be required to drive high diffusive fluxes, and processes other than simple diffusion may play a role in the rapid exchanges which do occur [82]. Given this uncertainty we conservatively restricted such transport to species known or expected to be exchanged between cell types (under at least some circumstances); a complete list is given in Outline.

Net import or export of metabolites from the system was limited to the mesophyll, for gases exchanged with the intercellular airspace, or the bundle sheath, for soluble metabolites exchanged with the leaf's vascular system. Reactions were not otherwise restricted *a priori* to a particular cell type. To facilitate integration with cell-type-specific RNA data, gene associations in this model are tagged with the relevant cell type, e.g. 'bs.GRMZM2G039273' vs 'ms.GRMZM2G039273'.

Leaf gradient model. The choice of phloem transport metabolites (other than sucrose) is a compromise. Glycine is the most abundant amino acid in maize phloem [83], and glutathione is a putative phloem sulfur transport compound [84], but many other amino acids are present in the phloem sap, and other compounds (e.g., S-methyl-methionine [84]) may play roles in phloem sulfur transport. However, we found that the available data did not adequately constrain rates of phloem transport if multiple transport species of each type were allowed, resulting in high rates of transport from the base towards the tip, against the direction of bulk flow in the phloem.

For simplicity, export of metabolites from the leaf to the rest of the plant through the phloem was neglected and net import of sucrose was not allowed. Each segment was taken to have the same total area, so that a $1 \mu\text{mol m}^{-2} \text{s}^{-1}$ rate of sucrose loading in one segment exactly balanced a $1 \mu\text{mol m}^{-2} \text{s}^{-1}$ rate of sucrose unloading in another segment.

Note that the whole-leaf model is constructed dynamically within the data-fitting code, rather than being loaded from an SBML file.

Physiological constraints. Rubisco carboxylase and oxygenase rates v_c and v_o in mesophyll and bundle sheath chloroplasts were constrained to obey Michaelis-Menten kinetic laws with competitive inhibition,

$$(5) \quad \begin{aligned} v_c &= \frac{v_{c,\max} [\text{CO}_2]}{[\text{CO}_2] + k_c \left(1 + \frac{[\text{O}_2]}{k_o}\right)} \\ v_o &= \frac{v_{o,\max} [\text{O}_2]}{[\text{O}_2] + k_o \left(1 + \frac{[\text{CO}_2]}{k_c}\right)}, \end{aligned}$$

and the relationship $v_{o,\max}/v_{c,\max} = k_C/(k_O \cdot S_R)$ was imposed, from which eq. (2) follows [15]. The Michaelis-Menten constants for oxygen and carbon dioxide k_C and k_O and the Rubisco specificity S_R were set to values typical of C4 species: k_C , 650 $\mu\text{mol mol}^{-1}$; k_O , 450 mmol mol^{-1} ; S_R , 2590 [15].

The rate of PEP carboxylation in the mesophyll was bounded above by an appropriate kinetic law,

$$(6) \quad v_p = \frac{v_{p,\max} [\text{CO}_2]}{k_{C,p} + [\text{CO}_2]}$$

with $0 \leq v_{p,\text{active}} \leq v_{p,\max}$ and an appropriate $k_{C,p}$ (80 mmol mol^{-1} , [15]).

The parameters $v_{p,\max}$ and $v_{c,\max}$ representing the total amount of Rubisco and PEPC available may be fixed to permit comparison to models parameterized in those terms or allowed to vary.

Rates of oxygen and carbon dioxide diffusion from the bundle sheath to the mesophyll, L and L_O , were constrained to obey the relationship

$$(7) \quad \begin{aligned} L &= g_{BS} (\text{CO}_{2,BS} - \text{CO}_{2,ME}) \\ L_O &= g_{BS,O} (\text{O}_{2,BS} - \text{O}_{2,ME}) \end{aligned}$$

with $g_{BS,O} = 0.047g_{BS}$ [15]. All simulations used the bundle sheath CO_2 conductivity measured by [35] for maize plants grown under high light, $1.03 \pm 0.18 \mu\text{mol m}^{-2} \text{s}^{-1}$. While g_{BS} undoubtedly varies along the developmental gradient, its deviation from this value (measured in fully-expanded leaves, 3-4 weeks after planting) is likely greatest below the region of high suberin synthesis identified 4 cm from the leaf base [30]; as the C4 cycle was not predicted to operate at high rates in this region, the impact of this discrepancy should be limited.

Resistance to CO_2 diffusion from the intercellular airspace to the mesophyll cells was neglected; ref. [85] reported $g_m \approx 1 \text{ mmol m}^{-2} \text{s}^{-1}$ in maize under a variety of conditions, suggesting the mesophyll and intercellular CO_2 levels would differ only slightly at the rates of CO_2 assimilation and release dealt with here. Similarly, all intracellular compartments were taken to have equal CO_2 concentrations.

Optimization calculations. The nonlinear modeling package uses the libsbml python bindings to read SBML files [86] and an internal representation of SBML models derived from the SloppyCell package [87, 88]. IPOPT calculations used version 3.11.8 with the linear solver ma97 from the HSL Mathematical Software Library [89]. Where not specified, convergence tolerance was 10^{-5} , or 10^{-4} in FVA calculations. To solve purely linear problems (e.g., to test the production of biomass species during the reconstruction process, where nonlinear constraints were not used) the GNU Linear Programming Kit, version 4.47 [90], was called through a Python interface [91].

Comparison with other models. Python code used to calculate the predictions of the models of von Caemmerer [15] for comparison with nonlinear optimization results is provided in S14 Protocol.

Integrating biochemical and RNA-seq data.

RNA-seq datasets. To obtain mesophyll- and bundle-sheath-specific expression levels at 15 points, we combined the non-tissue-type-specific data of Wang et al. [30], measured at 1-cm spatial resolution, with the tissue-specific data of Tausta et al. [31] obtained by using laser capture microdissection (LCM) – measured 4 cm, 8 cm and 13 cm from the leaf base (the upper three highlighted positions in Fig. 3b). This integration was achieved by determining for each gene at each of those points with LCM data the ratio of the average RPKM in the mesophyll (M) to the sum of the average RPKM values for mesophyll and bundle sheath ($M + B$); furthermore, we assumed that the $M/(M + B)$ ratio at the leaf base was 0.5 (based on the proteomic experiments of Majeran et al. [37], which showed only limited mesophyll-bundle sheath specificity there), and linearly interpolating to estimate $M/(M + B)$ ratios at all 15 points. For very weakly expressed genes, we did not impose cell-type specificity: where the sum of mesophyll and bundle sheath RPKM in the LCM data was less than 0.1, we assumed $M/(M + B) = 0.5$. We then divided the mean whole-leaf FPKM measurement at each point into mesophyll and bundle sheath portions according to these ratios.

To associate expression data with a reaction, data for its associated genes were summed, dividing the data for a gene associated with multiple reactions in the model equally among them. The uncertainties δ_{ij} in the objective function (eq. (3)) were estimated in an ad hoc way by splitting the standard deviations of the FPKM values over multiple experimental replicates according to the $M/(M + B)$ ratios and then summing the uncertainties for all genes associated with a particular reaction, imposing a minimum relative error of 0.05 and a minimum absolute uncertainty corresponding to 7.5 FPKM.

To globally rescale the expression data to be comparable to expected flux values, data for PEPC and Rubisco were compared to the enzyme activity measurements discussed below and a simple linear regression performed, yielding a conversion factor of $204 \text{ FPKM} \approx 1 \mu\text{mol m}^{-2} \text{ s}^{-1}$ for these enzymes. All expression data were divided by this factor before solving the optimization problem.

Enzyme activity measurements. Enzyme activities constrained by measurements in [30] were alanine aminotransferase, aspartate aminotransferase, fructose biphosphate aldolase, glyceraldehyde 3-phosphate dehydrogenase (NADPH), glyceraldehyde 3-phosphate dehydrogenase (NADH), glutamate dehydrogenase (NADH), malate dehydrogenase (NADH), malate dehydrogenase (NADPH), PEPC, phosphofructokinase, phosphoglucosmutase, phosphoglucose isomerase, phosphoglycerokinase, Rubisco, transketolase, triose phosphate isomerase, and UDP-glucose pyrophosphorylase.

For Rubisco and PEPC, enzyme data constrained the sum of the variable kinetic parameters $v_{c,\text{max}}$ and $v_{p,\text{max}}$ in mesophyll and bundle sheath compartments, rather than the sum of the associated fluxes. Enzyme data in nanomole per minute per gram fresh weight was converted to micromole per second per square meter of leaf surface area assuming a fresh weight of 150 g m^{-2} .

Handling reversible reactions. The objective function (eq. (3)) optimizes the agreement between the absolute value of the flux through each reaction with its data, but IPOPT requires a twice continuously differentiable objective function. We use a reformulation F' representing each absolute value $|v_{ij}|$ as the product of the flux and a parameter σ_{ij} representing its sign:

$$(8) \quad F'(v) = \sum_{i=0}^{N_r} \sum_{j=1}^{15} \frac{(e^{s_i} \sigma_{ij} v_{ij} - d_{ij})^2}{\delta_{ij}^2} + \alpha \sum_{i=0}^{N_r} s_i^2$$

Similarly, the enzyme activity data constraint, eq. (4), was rewritten to replace absolute values in this way. Reaction rates with positive (negative) sign parameter were required to take values greater than a small negative (less than a small positive) tolerance, typically 1.0.

Choosing the σ_{ij} to optimize F' is a very large scale mixed-integer nonlinear programming problem. We arrive at an approximate solution using a heuristic method similar in spirit to that of [32], with three steps.

- (1) The subproblems representing each segment of the leaf are solved separately, with all scales s_i set to zero and modest upper and lower bounds on the reactions representing nutrient exchange with the phloem. Within each segment, a sign for the reversible reaction r_1 with the highest associated expression data is chosen by first setting its sign σ_1 to +1, finding the minimum-flux best-fitting flux distribution \mathbf{v}^+ ignoring the costs associated with all other reversible reactions (but including costs associated with all irreversible reactions), then finding the cost c^+ of the best-fitting flux distribution \mathbf{v}^+ considering the costs of the reversible reactions with nonzero fluxes in \mathbf{v}^+ (temporarily setting their signs according to their values in that case.) A cost c^- is determined analogously after setting the sign σ_1 to -1 , and if $c^- < c^+$, $\sigma_1 = -1$ is chosen; otherwise, $\sigma_1 = +1$. Then the reversible reaction with the second-highest expression data r_2 is treated in the same way, considering r_1 to be irreversible.
- (2) When signs for all reversible reactions have been chosen at a segment, a final best-fitting flux distribution given those signs is determined. Then the full optimization problem, combining all fifteen segments, is solved with the chosen sign parameters fixed, using those flux distributions to provide a nearly-feasible initial guess.
- (3) The sign-choice process in each subproblem is then solved again, fixing the scale factors s_i and rates of metabolite exchange with the phloem to those determined in the full problem. If no signs change, or if the new signs do not decrease the objective function value, fitting stops; otherwise, step 2 is repeated.
- (4) Finally, for each reaction j with nonzero data and maximum absolute flux less than 0.0001 at any point in the leaf model, a lower bound of $-0.99d_i$ is imposed on the term $(e^{s_i} \sigma_{ij} v_{ij} - d_{ij})$ in the objective function, for $i = 1, \dots, 15$, and the full fifteen-segment optimization problem is solved again.

The final step addresses the observation that the optimization process occasionally converged to a solution in which a few reactions with associated data were predicted to have zero flux when a better solution with nonzero flux existed. In some cases

(e.g. the $s_i = 0$ case shown in Supporting Figure 4) this step did not lead to an overall reduction in the objective function and was omitted.

Steps 1 and 3 take between one and eight hours per segment using an AMD Opteron 6272 and may be easily parallelized across up to 15 processors. Step 2 may take up to 2 hours in the first iteration but is often faster in later iterations, when the initial guess is closer to the optimum. Typically the procedure stops after 4-5 iterations, requiring about 24 total hours of wall time using 15 processors.

Special cases. The Rubisco oxygenase, Rubisco carboxylase, and mesophyll PEPC fluxes are excluded from the objective function. Instead, terms are added comparing the transcriptomic data for those enzymes to the variables which explicitly represent their activity level: for Rubisco, $v_{c,\max}$ in mesophyll and bundle sheath compartments, and for PEPC, $v_{p,\max}$ in the mesophyll. Scale factors for the mesophyll and bundle sheath Rubisco activities are not constrained to be equal.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation grant IOS-1127017 and a grant to the International Rice Research Institute from the Bill and Melinda Gates Foundation. The authors thank Tom Brutnell, Lin Wang, Lori Tausta, Qi Sun, Zehong Ding, and Tim Nelson for data and comments, Sue Rhee, Kate Dreher, and Peifen Zhang for helpful discussion of our use of CornCyc, and Lei Huang and Brandon Barker for discussions of metabolic modeling.

REFERENCES

1. von Caemmerer S, Furbank RT (2003) The C(4) pathway: an efficient CO(2) pump. *Photosynthesis Research* 77: 191–207.
2. Sage RF, Christin PA, Edwards EJ (2011) The C4 plant lineages of planet Earth. *Journal of Experimental Botany* 62: 3155–3169.
3. Brown RH (1999) 14 - Agronomic implications of C4 photosynthesis. In: Monson RK, Sage RF, editors, *C4 Plant Biology*, San Diego: Academic Press, *Physiological Ecology*. pp. 473–507.
4. Kanai R, Edwards GE (1999) 3 - The Biochemistry of C4 Photosynthesis. In: Monson RKK, Sage, Rowan F, editors, *C4 Plant Biology*, San Diego: Academic Press, *Physiological Ecology*. pp. 49–87.
5. Hibberd J, Covshoff S (2010) The Regulation of Gene Expression Required for C-4 Photosynthesis. *Annual Review of Plant Biology* 61: 181–207.
6. Studer AJ, Gandin A, Kolbe AR, Wang L, Cousins AB, et al. (2014) A limited role for carbonic anhydrase in C4 photosynthesis as revealed by a calca2 double mutant in maize. *Plant Physiology* : pp.114.237602.
7. Furbank RT (2011) Evolution of the C4 photosynthetic mechanism: are there really three C4 acid decarboxylation types? *Journal of Experimental Botany* 62: 3103–3108.
8. Sage RF (2004) The evolution of C4 photosynthesis. *New Phytologist* 161: 341–370.
9. Christin PA, Samaritani E, Petitpierre B, Salamin N, Besnard G (2009) Evolutionary Insights on C4 Photosynthetic Subtypes in Grasses from Genomics and Phylogenetics. *Genome Biology and Evolution* 1: 221–230.
10. Griffiths H, Weller G, Toy LFM, Dennis RJ (2013) You're so vein: bundle sheath physiology, phylogeny and evolution in C3 and C4 plants. *Plant, Cell & Environment* 36: 249–261.
11. Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, et al. (2013) Predicting C4 Photosynthesis Evolution: Modular, Individually Adaptive Steps on a Mount Fuji Fitness Landscape. *Cell* 153: 1579–1588.

12. Way DA, Katul GG, Manzoni S, Vico G (2014) Increasing water use efficiency along the C3 to C4 evolutionary pathway: a stomatal optimization perspective. *Journal of Experimental Botany* : eru205.
13. Covshoff S, Hibberd JM (2012) Integrating C4 photosynthesis into C3 crops to increase yield potential. *Current Opinion in Biotechnology* 23: 209–214.
14. von Caemmerer S, Quick WP, Furbank RT (2012) The development of C4 rice: current progress and future challenges. *Science* 336: 1671–1672.
15. von Caemmerer S (2000) Biochemical models of leaf photosynthesis. Number 2 in *Techniques in plant sciences*. Collingwood, VIC: CSIRO Publishing.
16. Wang Y, Long SP, Zhu XG (2014) Elements Required for an Efficient NADP-ME Type C4 Photosynthesis. *Plant Physiology* : pp.113.230284.
17. Wang Y, Bräutigam A, Weber APM, Zhu XG (2014) Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. *Journal of Experimental Botany* 65: 3567–3578.
18. Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? *Nature Biotechnology* 28: 245–248.
19. Boyd S, Vandenberghe L (2004) *Convex Optimization*. Cambridge University Press.
20. de Oliveira Dal’Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis. *Plant Physiology* 152: 579–589.
21. Saha R, Suthers PF, Maranas CD (2011) Zea mays iRS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism. *PLoS ONE* 6: e21784.
22. Wächter A, Biegler LT (2006) On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming* 106: 25–57.
23. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19: 524–531.
24. Li P, Ponnala L, Gandotra N, Wang L, Si Y, et al. (2010) The developmental dynamics of the maize leaf transcriptome. *Nature Genetics* 42: 1060–1067.
25. Plant Metabolic Network (PMN) (2013). CornCyc 4.0. <http://pmn.plantcyc.org/CORN/organism-summary> on www.plantcyc.org.
26. Sun Q, Zybaïlov B, Majeran W, Friso G, Olinares PDB, et al. (2009) PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Research* 37: D969–974.
27. Reed J, Vo T, Schilling C, Palsson B (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biology* 4: R54.
28. Xu E (2011). Pyipopt. <http://github.com/xuy/pyipopt>.
29. Nelson T (2011) The grass leaf developmental gradient as a platform for a systems understanding of the anatomical specialization of C4 leaves. *Journal of Experimental Botany* 62: 3039–3048.
30. Wang L, Czedik-Eysenberg A, Mertz RA, Si Y, Tohge T, et al. (2014) Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice. *Nature Biotechnology* 32: 1158–1165.
31. Tausta SL, Li P, Si Y, Gandotra N, Liu P, et al. (2014) Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C4-related processes. *Journal of Experimental Botany* 65: 3543–3555.
32. Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, et al. (2012) Improving metabolic flux predictions using absolute gene expression data. *BMC Systems Biology* 6: 73.
33. Barker B, Sadagopan N, Wang Y, Smallbone K, Myers CR, et al. (2014) A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. [arXiv:1404.4755 \[q-bio\]](https://arxiv.org/abs/1404.4755).
34. Mahadevan R, Schilling C (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering* 5: 264–276.
35. Bellasio C, Griffiths H (2014) Acclimation to low light by C4 maize: implications for bundle sheath leakiness. *Plant, Cell & Environment* 37: 1046–1058.
36. Hatch MD (1987) C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics* 895: 81–106.

37. Majeran W, Friso G, Ponnala L, Connolly B, Huang M, et al. (2010) Structural and Metabolic Transitions of C4 Leaf Development and Differentiation Defined by Microscopy and Quantitative Proteomics in Maize. *The Plant Cell* 22: 3509–3542.
38. Wingler A, Walker RP, Chen ZH, Leegood RC (1999) Phosphoenolpyruvate Carboxykinase Is Involved in the Decarboxylation of Aspartate in the Bundle Sheath of Maize. *Plant Physiology* 120: 539–546.
39. Ponnala L, Wang Y, Sun Q, van Wijk KJ (2014) Correlation of mRNA and protein abundance in the developing maize leaf. *The Plant Journal* 78: 424–440.
40. Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, et al. (2009) Interpreting Expression Data with Metabolic Flux Models: Predicting *Mycobacterium tuberculosis* Mycolic Acid Production. *PLoS Comput Biol* 5: e1000489.
41. Gomes de Oliveira Dal’Molin C, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) C4GEM - Genome-Scale Metabolic Model to study C4 plant metabolism. *Plant Physiology* : pp.110.166488.
42. Simons M, Saha R, Amiour N, Kumar A, Guillard L, et al. (2014) Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model. *Plant Physiology* 166: 1659–1674.
43. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research* 42: D199–205.
44. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28: 27–30.
45. Plant Metabolic Network (PMN) (2014). Enzyme functional annotation method. http://www.plantcyc.org/about/databases_overview.faces#e2p2 on www.plantcyc.org. October 16, 2014.
46. Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* 9: e1003285.
47. Latendresse M, Krummenacker M, Trupp M, Karp PD (2012) Construction and completion of flux balance models from pathway databases. *Bioinformatics* 28: 388–396.
48. Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols* 5: 93–121.
49. Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics* 11: 40–79.
50. Plant Metabolic Network (PMN) (2014). Pmn database content statistics. http://www.plantcyc.org/release_notes/content_statistics.faces on www.plantcyc.org. January 13, 2015.
51. Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, et al. (2014) The Sol Genomics Network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Research* : gku1195.
52. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, et al. (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Research* 42: D1193–D1199.
53. Urbanczyk-Wochniak E, Sumner LW (2007) MedicCyc: a biochemical pathway database for *Medicago truncatula*. *Bioinformatics* 23: 1418–1423.
54. Naithani S, Raja R, Waddell EN, Elser J, Gouthu S, et al. (2014) VitisCyc: a metabolic pathway knowledgebase for grapevine (*Vitis vinifera*). *Frontiers in Plant Science* 5: 644.
55. Jung S, Ficklin SP, Lee T, Cheng CH, Blenda A, et al. (2014) The Genome Database for Rosaceae (GDR): year 10 update. *Nucleic Acids Research* 42: D1237–D1244.
56. Chang RL, Ghamsari L, Manichaikul A, Hom EFY, Balaji S, et al. (2011) Metabolic network reconstruction of *Chlamydomonas* offers insight into light-driven algal metabolism. *Molecular Systems Biology* 7: 518.
57. Mahadevan R, Edwards JS, Doyle FJ (2002) Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*. *Biophysical Journal* 83: 1331–1340.
58. Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing - bridging the gap between constraint-based and kinetic modelling. *FEBS Journal* 274: 5576–5585.

59. Jamshidi N, Palsson BØ (2010) Mass Action Stoichiometric Simulation Models: Incorporating Kinetics and Regulation into Stoichiometric Models. *Biophysical Journal* 98: 175–185.
60. Feng X, Xu Y, Chen Y, Tang YJ (2012) Integrating Flux Balance Analysis into Kinetic Models to Decipher the Dynamic Metabolism of *Shewanella oneidensis* MR-1. *PLoS Computational Biology* 8: e1002376.
61. Cotten C, Reed JL (2013) Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. *BMC Bioinformatics* 14: 32.
62. Chowdhury A, Zomorodi AR, Maranas CD (2014) k-OptForce: Integrating Kinetics with Flux Balance Analysis for Strain Design. *PLoS Computational Biology* 10: e1003487.
63. Tan Y, Lafontaine Rivera JG, Contador CA, Asenjo JA, Liao JC (2011) Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. *Metabolic Engineering* 13: 60–75.
64. Hoppe A (2012) What mRNA Abundances Can Tell us about Metabolism. *Metabolites* 2: 614–631.
65. Schwender J, König C, Klapperstück M, Heinzel N, Munz E, et al. (2014) Transcript abundance on its own cannot be used to infer fluxes in central metabolism. *Frontiers in Plant Science* .
66. Machado D, Herrgård M (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLoS Computational Biology* 10: e1003580.
67. Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nature Biotechnology* 28: 1279–1285.
68. Salimi F, Zhuang K, Mahadevan R (2010) Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing. *Biotechnology Journal* 5: 726–738.
69. Zhuang K, Izallalen M, Mouser P, Richter H, Risso C, et al. (2011) Genome-scale dynamic modeling of the competition between *Rhodospirillum rubrum* and *Geobacter* in anoxic subsurface environments. *The ISME Journal* 5: 305–316.
70. Zomorodi AR, Maranas CD (2012) OptCom: A Multi-Level Optimization Framework for the Metabolic Modeling and Analysis of Microbial Communities. *PLoS Comput Biol* 8: e1002363.
71. Zengler K, Palsson BO (2012) A road map for the development of community systems (CoSy) biology. *Nature Reviews Microbiology* 10: 366–372.
72. Khandelwal RA, Olivier BG, Röling WFM, Teusink B, Bruggeman FJ (2013) Community Flux Balance Analysis for Microbial Consortia at Balanced Growth. *PLoS ONE* 8: e64567.
73. Chiu HC, Levy R, Borenstein E (2014) Emergent Biosynthetic Capacity in Simple Microbial Communities. *PLoS Computational Biology* 10: e1003695.
74. Zomorodi AR, Islam MM, Maranas CD (2014) d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synthetic Biology* 3: 247–257.
75. Stolyar S, Dien SV, Hillesland KL, Pinel N, Lie TJ, et al. (2007) Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology* 3: 92.
76. Bordbar A, Lewis NE, Schellenberger J, Palsson BØ, Jamshidi N (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Molecular Systems Biology* 6: 422.
77. Heinken A, Sahoo S, Fleming RMT, Thiele I (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes* 4: 28–40.
78. Cheung CYM, Poolman MG, Fell DA, Ratcliffe RG, Sweetlove LJ (2014) A Diel Flux Balance Model Captures Interactions between Light and Dark Metabolism during Day-Night Cycles in *C3* and *Crassulacean Acid Metabolism* Leaves. *Plant Physiology* 165: 917–929.
79. Grafahrend-Belau E, Junker A, Eschenröder A, Müller J, Schreiber F, et al. (2013) Multi-scale Metabolic Modeling: Dynamic Flux Balance Analysis on a Whole-Plant Scale. *Plant Physiology* 163: 637–647.
80. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols* 2: 727–738.

81. Weiner H, Burnell JN, Woodrow IE, Heldt HW, Hatch MD (1988) Metabolite Diffusion into Bundle Sheath Cells from C4 Plants Relation to C4 Photosynthesis and Plasmodesmatal Function. *Plant Physiology* 88: 815–822.
82. Sowiński P, Szczepanik J, Minchin PEH (2008) On the mechanism of C4 photosynthesis intermediate exchange between Kranz mesophyll and bundle sheath cells in grasses. *Journal of Experimental Botany* 59: 1137–1147.
83. Ohshima T, Hayashi H, Chino M (1990) Collection and Chemical Composition of Pure Phloem Sap from *Zea mays* L. *Plant and Cell Physiology* 31: 735–737.
84. Bourgis F, Roje S, Nuccio ML, Fisher DB, Tarczynski MC, et al. (1999) S-Methylmethionine Plays a Major Role in Phloem Sulfur Transport and Is Synthesized by a Novel Type of Methyltransferase. *The Plant Cell* 11: 1485–1497.
85. Kromdijk J, Griffiths H, Schepers HE (2010) Can the progressive increase of C4 bundle sheath leakiness at low PFD be explained by incomplete suppression of photorespiration? *Plant, Cell & Environment* 33: 1935–1948.
86. Bornstein BJ, Keating SM, Jouraku A, Hucka M (2008) LibSBML: an API library for SBML. *Bioinformatics* 24: 880–881.
87. Gutenkunst RN, Atlas JC, Casey FP, Daniels BC, Kuczenski RS, et al. (2007). SloppyCell. <http://sloppyCell.sourceforge.net>.
88. Myers C, Gutenkunst R, Sethna J (2007) Python Unleashed on Systems Biology. *Computing in Science and Engineering* 9: 34–37.
89. HSL (2013). A collection of fortran codes for large scale scientific computation. <http://www.hsl.rl.ac.uk>.
90. GLPK (2011). Gnu linear programming kit, version 4.47. <http://www.gnu.org/software/glpk/glpk.html>.
91. Finley T (2008). pyglpk. <http://tfinley.net/software/pyglpk>.
92. SRI International, Menlo Park, CA (2013) Pathway Tools v. 17.0 user manual.
93. Elthon TE, Stewart CR (1982) Proline Oxidation in Corn Mitochondria Involvement of NAD, Relationship to Ornithine Metabolism, and Sidedness on the Inner Membrane. *Plant Physiology* 70: 567–572.
94. Brownleader M, Harborne J, Dey P (1997) Carbohydrate metabolism: primary metabolism of polysaccharides. In: Dey P, Harborne J, editors, *Plant biochemistry*, San Diego: Academic Press. pp. 111–142.
95. Allen JF (2003) Cyclic, pseudocyclic and noncyclic photophosphorylation: new links in the chain. *Trends in Plant Science* 8: 15–19.
96. Asada K (1999) The Water-Water Cycle in Chloroplasts: Scavenging of Active Oxygens and Dissipation of Excess Photons. *Annual Review of Plant Physiology and Plant Molecular Biology* 50: 601–639.
97. Foyer C, Harbinson J (1997) The photosynthetic electron transport system: efficiency and control. In: Foyer C, Quick W, editors, *A Molecular Approach to Primary Metabolism in Higher Plants*, London: Taylor and Francis.
98. Fettke J, Hejazi M, Smirnova J, Höchel E, Stage M, et al. (2009) Eukaryotic starch degradation: integration of plastidial and cytosolic pathways. *Journal of Experimental Botany* 60: 2907–2922.
99. Streb S, Zeeman SC (2012) Starch Metabolism in Arabidopsis. *The Arabidopsis Book* : e0160.
100. Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, et al. (2013) Acyl-Lipid Metabolism. *The Arabidopsis Book* : e0161.
101. Ohlrogge J, Browse J (1995) Lipid biosynthesis. *The Plant Cell* 7: 957–970.
102. Dörmann P, Benning C (1998) The role of UDP-glucose epimerase in carbohydrate metabolism of Arabidopsis. *The Plant Journal* 13: 641–652.
103. BRENDA (2013). Information on EC 5.1.3.2 - UDP-glucose 4-epimerase. <http://brenda-enzymes.org/enzyme.php?ecno=5.1.3.2>. October 9, 2013.
104. Plant Metabolic Network (PMN) (2014). *Arabidopsis thaliana* col pathway: palmitoleate biosynthesis II. <http://pmn.plantcyc.org/ARA/NEW-IMAGE?type=PATHWAY&object=PWY-5366> on www.plantcyc.org. October 16, 2014.
105. Gibson KJ (1993) Palmitoleate formation by soybean stearyl-acyl carrier protein desaturase. *Biochimica Et Biophysica Acta* 1169: 231–235.

106. Sperling P, Schmidt H, Heinz E (1995) A Cytochrome-b5-Containing Fusion Protein Similar to Plant Acyl Lipid Desaturases. *European Journal of Biochemistry* 232: 798–805.
107. Harwood JL (1996) Recent advances in the biosynthesis of plant fatty acids. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* 1301: 7–56.
108. Shanklin J, Cahoon EB (1998) Desaturation and Related Modifications of Fatty Acids. *Annual Review of Plant Physiology and Plant Molecular Biology* 49: 611–641.
109. Poolman MG, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of arabidopsis and some of its properties. *Plant Physiol* 151: 1570–1581.
110. Poolman MG, Kundu S, Shaw R, Fell DA (2013) Responses to Light Intensity in a Genome-Scale Model of Rice Metabolism. *Plant Physiology* 162: 1060–1072.
111. Mintz-Oron S, Meir S, Malitsky S, Ruppin E, Aharoni A, et al. (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences* 109: 339–344.
112. Reumann S, Weber APM (2006) Plant peroxisomes respire in the light: Some gaps of the photorespiratory C2 cycle have become filled—Others remain. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1763: 1496–1510.
113. Foyer CH, Bloom AJ, Queval G, Noctor G (2009) Photorespiratory Metabolism: Genes, Mutants, Energetics, and Redox Signaling. *Annual Review of Plant Biology* 60: 455–484.
114. Weber AP, Linka N (2011) Connecting the Plastid: Transporters of the Plastid Envelope and Their Role in Linking Plastidial with Cytosolic Metabolism. *Annual Review of Plant Biology* 62: 53–77.
115. Bräutigam A, Weber APM (2011) Chapter 11 Transport Processes: Connecting the Reactions of C4 Photosynthesis. In: Raghavendra AS, Sage RF, editors, *C4 Photosynthesis and Related CO2 Concentrating Mechanisms*, Springer Netherlands, number 32 in *Advances in Photosynthesis and Respiration*. pp. 199–219. URL http://link.springer.com/chapter/10.1007/978-90-481-9407-0_11.
116. Douce R, Heldt HW (2000) Photorespiration. In: Leegood R, Sharkey T, von Caemmerer S, editors, *Photosynthesis: Physiology and Metabolism*, Boston: Kluwer Academic Publishers.
117. Hanson AD, Roje S (2001) One-carbon metabolism in higher plants. *Annual Review of Plant Physiology and Plant Molecular Biology* 52: 119–137.
118. Bartoli CG, Pastori GM, Foyer CH (2000) Ascorbate Biosynthesis in Mitochondria Is Linked to the Electron Transport Chain between Complexes III and IV. *Plant Physiology* 123: 335–344.
119. Foyer C, Rowell J, Walker D (1983) Measurement of the ascorbate content of spinach leaf protoplasts and chloroplasts during illumination. *Planta* 157: 239–244.
120. Smirnoff N (1996) The Function and Metabolism of Ascorbic Acid in Plants. *Annals of Botany* 78: 661–669.
121. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* 300: 1005–1016.
122. Munekage Y, Hashimoto M, Miyake C, Tomizawa KI, Endo T, et al. (2004) Cyclic electron flow around photosystem I is essential for photosynthesis. *Nature* 429: 579–582.
123. Shikanai T (2007) Cyclic Electron Transport Around Photosystem I: Genetic Approaches. *Annual Review of Plant Biology* 58: 199–217.
124. Takabayashi A, Kishine M, Asada K, Endo T, Sato F (2005) Differential use of two cyclic electron flows around photosystem I for driving CO2-concentration mechanism in C4 photosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 102: 16898–16903.
125. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Research* 32: W327–W331.
126. Lopez-Calcagno PE, Howard TP, Raines CA (2014) The CP12 protein family: a thioredoxin-mediated metabolic switch? *Frontiers in Plant Science* 5.
127. Rizov I, Doulis A (2000) Determination of glycerolipid composition of rice and maize tissues using solid-phase extraction. *Biochemical Society Transactions* 28: 586.
128. Leech RM, Rumsby MG, Thomson WW (1973) Plastid Differentiation, Acyl Lipid, and Fatty Acid Changes in Developing Green Maize Leaves. *Plant Physiology* 52: 240–245.

129. Plant Metabolic Network (PMN) (2014). *Zea mays* mays pathway: homogalacturonan biosynthesis. <http://pmn.plantcyc.org/CORN/NEW-IMAGE?type=PATHWAY&object=PWY-1061> on www.plantcyc.org. October 16, 2014.
130. Herold A, Lewis DH (1977) Mannose and Green Plants: Occurrence, Physiology and Metabolism, and Use as a Tool to Study the Role of Orthophosphate. *New Phytologist* 79: 1–40.
131. Schnarrenberger C (1990) Characterization and compartmentation, in green leaves, of hexokinases with different specificities for glucose, fructose, and mannose and for nucleoside triphosphates. *Planta* 181: 249–255.
132. Plant Metabolic Network (PMN) (2014). PlantCyc pathway: D-mannose degradation. <http://pmn.plantcyc.org/PLANT/new-image?object=MANNCAT-PWY> on www.plantcyc.org. October 16, 2014.
133. Franceschi VR, Loewus FA (1995) Oxalate function and biosynthesis in plants and fungi. In: Khan SR, editor, *Calcium oxalate in biological systems*, CRC Press.
134. Franceschi VR, Nakata PA (2005) Calcium Oxalate in Plants: Formation and Function. *Annual Review of Plant Biology* 56: 41–71.
135. Debolt S, Melino V, Ford CM (2007) Ascorbate as a Biosynthetic Precursor in Plants. *Annals of Botany* 99: 3–8.
136. Lane BG, Dunwell JM, Ray JA, Schmitt MR, Cuming AC (1993) Germin, a protein marker of early plant development, is an oxalate oxidase. *Journal of Biological Chemistry* 268: 12239–12242.
137. Plant Metabolic Network (PMN) (2014). PlantCyc Reaction: 3.7.1.1. <http://pmn.plantcyc.org/PLANT/NEW-IMAGE?type=REACTION-IN-PATHWAY&object=OXALOACETASE-RXN> on www.plantcyc.org. October 16, 2014.
138. Hayaishi O, Shimazono H, Katagiri M, Saito Y (1956) Enzymatic formation of oxalate and acetate from oxaloacetate. *Journal of the American Chemical Society* 78: 5126–5127.
139. Burgener M, Suter M, Jones S, Brunold C (1998) Cyst(e)ine Is the Transport Metabolite of Assimilated Sulfur from Bundle-Sheath to Mesophyll Cells in Maize Leaves. *Plant Physiology* 116: 1315–1322.

LABORATORY OF ATOMIC AND SOLID STATE PHYSICS/INSTITUTE OF BIOTECHNOLOGY
CORNELL UNIVERSITY
ITHACA, NY
c.myers@cornell.edu

SUPPORTING INFORMATION

1. OUTLINE

This appendix describes the of creation of a metabolic model for maize from CornCyc. It covers the creation of an SBML model with exchange and biomass reactions and limited subcellular compartmentalization which can successfully simulate the production of many biomass components and photosynthetic carbon dioxide assimilation, the adaptation of the biomass equation from iRS1563, some considerations in the process of expanding the model to describe interacting mesophyll and bundle sheath compartments, and some modifications made in response to preliminary fitting results.

Sections 2 through 8 explain in detail the process of constructing the underlying metabolic model at the one-cell level. Section 9 discusses in detail changes made to gene associations based on early data fitting results. Section 10 describes changes to the iRS1563 biomass equation. Section 11 discusses plasmodesmatal transport in the two-cell model. Filenames referred to are in the `model.development` subdirectory of the project source code (see S15 Protocol.)

2. EXPORTING THE CORNCYC FBA MODEL FROM PATHWAY TOOLS

CornCyc 4.0 [25] was obtained from the Plant Metabolic Network and upgraded from from Pathway Tools 16.5 to 17.0 locally.

The frame `PWY-561` was removed from the database because otherwise some of the reactions of that pathway were excluded from the FBA export, apparently due to a bug.

A simple FBA problem was solved using the Pathway Tools FBA functionality [92], producing an output file which includes all reactions in the FBA model Pathway Tools generates internally, both those which are active in the solution to the FBA problem and those which are not. Note that this list of reactions is distinct from the list of reactions in the database itself; the Pathway Tools software prepares this set of reactions through an extensive process of excluding reactions which are unbalanced or otherwise undesirable while expanding reactions with classes of compounds as products or reactants into sets of possible specific instantiations which respect conservation of mass [47]. Working with the Pathway Tools FBA reaction set (rather than, e.g, an SBML export of the CornCyc database) allows us take advantage of this pre-processing; however, it comes at the cost of needing to reintroduce into the FBA model many reactions which are present in the CornCyc database but are excluded from the FBA export for one reason or another.

Reaction data was extracted from the FBA output file, and reactions were translated to refer to species by their CornCyc frame ID (to allow easy reference to the database and comparison with previous work, and avoid possible ambiguities.) Reactions were then added and removed from the model as described below.

3. DISCARDING REACTIONS

3.1. Polymerization reactions. Pathway Tools attempts to include an expanded representation of certain polymerization reactions in the exported FBA model, but this function is considered experimental [92]; these reactions were ignored. Note that some reactions representing polymer growth were added manually later in the process.

3.2. ATPases. We removed all reactions from CornCyc which have the effective stoichiometry

$$\begin{aligned} &\{ \text{'ADP'}: 1.0, \text{'ATP'}: -1.0, \text{'PROTON'}: 1.0, \text{'WATER'}: -1.0, \\ &\text{'|Pi|'}: 1.0 \} \end{aligned}$$

There are nine such reactions:

- RXN-11109,
- 3.6.4.6-RXN,
- RXN-11135,
- RXNO-1061,
- ADENOSINETRIPHOSPHATASE-RXN,
- 3.6.4.4-RXN,
- 3.6.4.9-RXN,
- 3.6.4.5-RXN,
- 3.6.4.3-RXN

all treated as reversible by the Pathway Tools export procedure. Typically these are simplified representations of the metabolic effect of enzymes whose complete function is outside the scope of the database, as, for example, EC 3.6.4.3, the microtubule-severing ATPase.

In their place, we added a single generic ATPase reaction to represent cellular maintenance costs, etc., with no associated genes.

3.3. Reactions involving generic electron donors and acceptors. Numerous reactions in the database are written with generic representations of electron carrier species ('a reduced electron acceptor', 'an oxidized electron acceptor'). Most of these reactions are outside the areas of emphasis of the model (e.g., brassinosteroid biosynthesis), have no curated pathway assignment, or also appear in forms which do specify the electron carrier species (e.g., the generic nitrate reductase reaction, NITRATEREDUCT-RXN, vs NITRATE-REDUCTASE-NADH-RXN,) and so could be safely neglected. A small set of exceptions identified in early drafts included reactions of fatty acid synthesis, handled as discussed below, and proline dehydrogenase, RXN-821, catalyzed by a mitochondrial-membrane-bound flavoprotein which donates electrons directly to the mitochondrial electron transport chain [93]. Because we have not thoroughly compartmentalized amino acid metabolism, we implemented this reaction as donating electrons to NAD^+ instead.

3.4. Duplicates. A number of other reactions were removed because they appeared to be exact (possibly unintentional) duplicates, down to gene associations, of other reactions in the database; or because they were being replaced by modified forms as discussed below. These are given in `reactions_to_remove.txt`.

3.5. Non-metabolic reactions. A number of reactions present in CornCyc were removed because the database indicated, e.g. through the Enzyme Commission summary for the relevant EC number, that they were primarily involved in extrametabolic functions (e.g., cell movement, regulation). These included the GTPases RXN-5462, 3.6.5.2-RXN, and 3.6.5.5-RXN.

3.6. Glucose-6-phosphate. In the reduced model (discussed below) only one reaction, myo-inositol-1-phosphate synthase, consumes the generic glucose-6-phosphate

species, rather than alpha-G6P or beta-G6P. To ensure that this reaction was appropriately connected to other G6P producing and consuming reactions we manually split it into two instances, one for alpha-G6P and one for beta-G6P.

3.7. UDP-glucose. For apparently all reactions in CornCyc involving UDP-glucose, the instantiation procedure produced one version involving generic UDP-D-glucose and one version involving UDP-alpha-D-glucose, the only child of the UDP-D-glucose class. UDP-alpha-D-glucose participated in almost no reactions other than these instantiations (in the reduced model, described below, only one: UDP-sulfoquinovose synthase, EC 3.13.1.1). As such there is little to distinguish the generic and specific versions of the reactions, which add complexity to the model and degeneracy to optimization predictions without providing significant information about the function of the system, so we removed the specific versions and changed the UDP-sulfoquinovose synthase to act on a generic UDP-D-glucose substrate.

4. MINOR REVISIONS TO ACHIEVE BASIC FUNCTIONALITY

4.1. Mitochondrial electron transport chain. The CornCyc representation of the mitochondrial electron transport pathway (PWY-3781, plus the mitochondrial ATPase (ATPSYN-RXN, EC 3.6.3.14)) was adjusted. Some reactions excluded from the initial Pathway Tools export because the balance state of reactions involving cytochrome C could not be determined were readadded manually; ubiquinones/ubiquinol were uniformly represented as ubiquinone-8/ubiquinol-8, and compartments were assigned to reactants and products to properly represent the transport of protons between the mitochondrial matrix and the mitochondrial intermembrane space. In CornCyc, as in MetaCyc and other related databases, transport of protons across the membrane is represented explicitly for complex I but not for complex III and complex IV; in agreement with the standard description of mitochondrial electron transport (see, e.g., [94]) proton transport was added to these reactions with a stoichiometry of 2 H⁺/e⁻ for complex III and 1 H⁺/e⁻ for complex IV. The stoichiometry of complex IV was further adjusted to include the H⁺ from the mitochondrial matrix that binds to oxygen to form water.

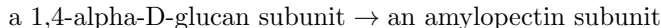
4.2. Photosynthesis: light reactions. Similarly, some modifications were made to the light reactions of photosynthesis (PWY-101). Reactions involving plastocyanins were not exported and were added manually; a chloroplastic ATP synthase and a reaction describing cyclic electron transport around PS I were added; and the stoichiometry of proton transport was adjusted in accordance with recent literature, assuming a Q cycle and ratio of 14 H⁺/3 ATP for the chloroplast ATP synthase [95].

Reduction of oxygen to superoxide at photosystem I (the Mehler reaction) was added to allow flux through the pathways of chloroplastic reactive oxygen species detoxification: superoxide dismutase and the ascorbate-glutathione cycle, including a reaction representing the direct, non-enzymatic reduction of monodehydroascorbate by ferredoxin [96,97].

4.3. Key reactions in biomass component production and nutrient uptake. Several components of biomass required either manual adjustment of reactions from the database or the addition of abstract synthesis reactions summarizing the behavior of pathways which could not easily be represented in more detail.

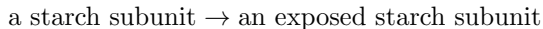
4.3.1. *Starch*. Starch synthase (GLYCOGENSYN-RXN) is not exported from CornCyc by default (it is a polymerization reaction, and marked as unbalanced in the PGDB); it was added manually in a form that produces the equivalent of one 1,4-alpha-D-glucan subunit.

The starch branching enzyme EC 2.4.1.18 (RXN-7710) is not exported from CornCyc by default (one reactant, starch, has an unspecified structure); it was added manually as



Note that this stoichiometry is not intended to suggest that the branching enzyme introduces branches at each subunit.

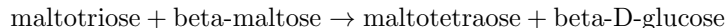
CornCyc provides a detailed reconstruction of the reactions of starch degradation (PWY-6724) which is by nature difficult to convert to a form suitable for FBA calculations, as many of the stoichiometry coefficients are undefined. To incorporate the effects of the glucan-water and phosphoglucan-water dikinases, for example, we would need to specify how many glucosyl residues must be phosphorylated (and then dephosphorylated) to produce “an exposed unphosphorylated, unbranched malto-oligosaccharide tail on amylopectin” of a given length; modeling the release of maltose from that tail would require an estimate of the typical unbranched length of such tails, etc. Rather than estimate average values for these parameters, we divide the reactions of the pathway into two types: those which condition starch for depolymerization, and actual depolymerization reactions. The first class (the dikinases above plus isoamylase) share the abstract stoichiometry



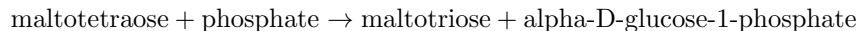
(neglecting any ATP costs), while the second class (beta amylase and disproportionating enzyme) convert exposed starch subunits to sugars appropriately.

The beta-maltose releasing reactions of the starch degradation pathway in CornCyc have no associated genes. We temporarily associated these reactions with the beta amylase record in the database (RXN-1827, EC 3.2.1.2) pending further review.

During transient starch degradation, beta-maltose and glucose are exported into the cytosol, where maltose is split, releasing one glucose molecule and donating one glucosyl residue to a cytosolic heteroglycan, from which it may be released in turn as glucose-1-phosphate [98]. In Arabidopsis, specific enzymes (DPE2 and PHS2) are known to be implicated in this process [99]. In simulations with this CornCyc-based FBA model we find the typical mode of breakdown of cytosolic maltose is to alpha-D-glucose and alpha-D-glucose-1-phosphate via AMYLOMALT-RXN,



and RXN0-5182,



effectively the standard pathway but with maltotriose/maltotetraose playing the role of the cytosolic heteroglycan pool. This approximation leads to a reasonable effective stoichiometry but it is possible that the genes associated with these reactions do not accurately represent the genes involved in the true underlying process; we have not systematically looked for maize counterparts of the Arabidopsis genes, for example.

4.3.2. *Cellulose*. The UDP-forming cellulose synthase, EC 2.4.1.12, is not exported from CornCyc by default (it is a polymerization reaction, and marked as unbalanced in the PGDB); it was added manually in a form that produces the equivalent of one subunit.

4.3.3. *Hemicellulose*. Similarly, the following hemicellulose polymerization reactions were added manually:

- 1,4-beta-D-xylan synthase, EC 2.4.2.24,
- reactions RXN-9093 (EC 2.4.2.-) and RXN-9094 (EC 2.4.1.-), representing the addition of arabinose and glucuronate to xylan to form arabinoxylan and glucuronoxylan respectively (note that the corresponding subunits notionally consist of one xylan subunit plus arabionose/glucuronate),
- glucomannan synthase, EC 2.4.1.32,
- RXN-9461 (EC 2.4.2.39), representing the addition of xylose to a glucan (as implemented, cellulose) to form xyloglucan (again, the corresponding effective subunit corresponds to one glucan subunit plus xylose)— note this representation ignores the previous step in CornCyc’s xyloglucan biosynthesis pathway, xyloglycan 4-glucosyltransferase (EC 2.4.1.168).

In addition to these explicit descriptions of hemicellulose formation from CornCyc, we added generic reactions representing the donation of the following sugar residues from activated donor molecules to unspecified generic polysaccharides:

- arabinose (from UDP-L-arabinose)
- galactose (from GDP-L-galactose)
- galacturonate (from UDP-D-galacturonate)
- glucose (from UDP-glucose)
- glucuronate (from UDP-D-glucuronate)
- mannose (from GDP-alpha-D-mannose)
- xylose (from UDP-alpha-D-xylose)

These reactions allow the model to represent flux of these sugars towards hemicelluloses or other polysaccharides without explicit synthesis pathways in CornCyc, or the construction of a hemicellulose term in the biomass equation in terms of the overall composition of hemicellulose without reference to specific synthesis reactions, as in our adaptation of the biomass reaction of iRS1563 (see the biomass reaction discussion, below.)

4.3.4. *Miscellaneous cell wall components*. The following additional cell wall component production reactions from CornCyc were added manually:

- 2.4.1.43-RXN, representing the formation of homogalacturonan from galacturonate
- RXN-9589 (EC 2.4.2.41), representing the addition of xylose to homogalacturonan to form xylogalacturonan (note the resulting xylogalacturonan subunit notionally consists of one galacturonate plus xylose)
- 13-BETA-GLUCAN-SYNTHASE-RXN (EC 2.4.1.12), representing the formation of callose from glucose.

Suberin production is not represented in CornCyc in detail but pathways for the synthesis of three key precursors, N-feruloyltyramine, octadecenedioate, and docosanediote, are provided. Sinks for N-feruloyltyramine and octadecenedioate

were added to the model to represent the flow of material towards suberin production; docosanedioate was neglected because no genes are associated with the reactions of its synthesis pathway. N-feruloyltyramine may be produced from trans-caffeate via either ferulate or caffeoyl-CoA; the branch through ferulate was initially dropped from the reduced version of the model used for data analysis because it relies on trans-feruloyl-CoA synthase, EC 6.2.1.34, which has no associated genes, but it was preserved in subsequent versions of the model because high expression levels for caffeate O-methyltransferase suggest this branch is indeed active.

(In CornCyc, the tyramine N-feruloyltransferase that produces N-feruloyltyramine from feruloyl-CoA could also catalyze the production of other hydroxycinnamic acid tyramine amides (cinnamoyltyramide, sinapoyltyramide, p-coumaroyl-tyramine) but we have neglected these for now.)

4.3.5. Fatty acids and lipids. Plant fatty acid and lipid biosynthesis is rich in complexity (see, e.g., [100]), and attempting to describe it in the FBA model at the level of detail at which it is currently understood would require a daunting number of reactions among the species representing the combinations of lipid head groups and acyl chains. Though CornCyc presents some pathways of lipid metabolism at such a high resolution, we have adopted a simplified approach which aims to include enough detail to allow the model to:

- predict based on RNA-seq data the total flow of biomass into fatty acids and lipids
- coarsely predict differences in the types of lipids and fatty acids produced, based on RNA-seq data
- approximately preserve the iRS1563 biomass equation.

The model describes in detail the sequence of reactions by which fatty acids up to lengths of 16 and 18 are synthesized in the chloroplast (though currently these reactions occur in the cytoplasmic compartment!), and the formation of oleate (as oleoyl-ACP) by the stearyl-ACP desaturase (PWY-5156; [100, 101]). In practice, these fatty acids may then enter the ‘prokaryotic’ pathway of glycerolipid synthesis in the chloroplast or leave the chloroplast and enter the ‘eukaryotic’ pathway of glycerolipid synthesis in the endoplasmic reticulum, with further desaturation of the acyl chains occurring after their incorporation into lipids.

We simplify this process by effectively decoupling the synthesis of different types of lipids (as distinguished by head groups) from the desaturation of their associated acyl chains. Reactions from lipid synthesis pathways are implemented as if all lipid species had one 16:0 and one 18:1 acyl chain, by implementing the glycerol-3-phosphate O-acyltransferase and 1-acylglycerol-3-phosphate O-acyltransferase reactions (RXN-10462 and 1-ACYLGLYCEROL-3-P-ACYLTRANSFER-RXN), written in the database with generic acyl-ACP substrates, with oleoyl-ACP and palmitoyl-ACP as substrates respectively. (This corresponds to the prokaryotic pathway; in the eukaryotic pathway oleoyl-CoA and palmitoyl-CoA would supply the acyl groups for diacylglycerol formation instead [100]. However the same genes are associated with the reactions of diacylglycerol synthesis in the two pathways (PWY-5667; PWY0-1319) in CornCyc and so they cannot be distinguished based on expression data alone; we have chosen one arbitrarily.)

This supply of diacylglycerol is sufficient to allow, without further modification to the CornCyc FBA export, the synthesis of a variety of lipids, including:

- phosphatidylcholine, phosphatidylethanolamine, phosphatidylglycerol, phosphatidylinositol;
- sulfoquinovosyldiacylglycerol.

UDP-glucose epimerase is exported from CornCyc in the UDP-glucose-producing direction by default; we allowed it to run in the reverse direction as well, consistent with literature evidence [102, 103], which allowed the production of mono- and digalactosyldiacylglycerol.

In sphingolipid metabolism, dihydrosphingosine, 4-hydroxysphinganine and sphinganine 1-phosphate may be produced, and sink reactions were added for them. Production of the ceramides and their derivatives would require the choice of a particular fatty acid source for the sphinganine acyltransferase, written by default with the generic substrate ‘a long-chain acyl-coA’; per the CornCyc description page for PWY-5129, in leaf sphingolipids C20 to C26 fatty acids are typical. Currently, the FBA model lacks a detailed implementation of production of very long chain fatty acids by elongation (a generic representation is present in CornCyc), so no supply of C20-26 fatty acids is available. We have deferred this issue to future work.

Separately, we model the desaturation of oleate to linoleate and linolenate and palmitate to palmitoleate. These (along with palmitate and stearate) are the fatty acid components of the iRS1563 biomass reaction, which originally incorporated them as triglycerides; our modified biomass equation consumes free fatty acids, rather than attempt to specify the precise ratios in which they are to be found in different lipid species in the leaf.

The CornCyc pathways for linoleate and linolenate produce them as lipid linoleoyl groups and lipid linolenoyl groups respectively, incorporated in generic lipid molecules; to allow these reactions to balance, and to provide linoleate and linolenate for the biomass reaction, we added lipases which release free linoleate/linolenate from the lipid linoleoyl and lipid linolenoyl groups, regenerating the pool of generic ‘lipid’ species (which participate only in the linoleate pathway, within the FBA model.) Note, however, that other reactions within the model but outside the indicated synthesis pathways are capable of producing linoleate and linolenate as well.

CornCyc includes no complete pathway for the production of palmitoleic acid; as there is experimental evidence it is produced in maize leaves (see the discussion of the biomass equation, below) we introduced the acyl-ACP $\Delta 9$ -desaturase reaction from the palmitoleate biosynthesis pathway of AraCyc (RXN-8389, 1.14.99.-), producing palmitoleoyl-ACP from palmitoyl-ACP [104], which restores this functionality (in combination with the palmitoleoyl-ACP hydrolase, RXN-9550, which is present in CornCyc.) Note that there is some evidence that the stearyl-ACP desaturase enzyme may also catalyze this reaction [105].

The oleoyl-acyl carrier protein hydrolase (EC 3.1.2.14) from CornCyc is unbalanced with respect to hydrogen; a version with an additional proton on the right hand side was added manually.

The $\Delta 9$ -desaturase and the desaturases producing linoleate and linolenate (RXN-9667 and RXN-9669) were written originally with generic electron donor and acceptor species. Initial review of the extensive literature on plant fatty acid desaturation suggests that the electron source for desaturases depends on their location within the cell, with chloroplastic desaturases accepting electrons from ferredoxin while desaturases in the endoplasmic reticulum accept electrons from NADH via cytochrome b5 or fused cytochrome domains (see, eg, [106–108].) As discriminating between

chloroplastic and extrachloroplastic fatty acid desaturation is not a high priority for the model, NADH was used as the sole electron donor for all three of these reactions.

The ferredoxin-dependent stearyl-ACP desaturase **RXN-7903**, not exported from the database by default because it is marked as unbalanced, was added in a form adjusted for hydrogen and charge balance. Ferredoxin-NADP oxidoreductase was made reversible to ensure NADPH can drive this reaction in the dark, as is observed [108].

4.3.6. Nucleic acids polymerization. Reactions representing the pyrophosphate-releasing incorporation of (d)NTPs into RNA and DNA were added and associated with the DNA-directed DNA polymerase and DNA-directed RNA polymerase reactions in the database. (In each case, it is assumed that all nucleotides occur with equal frequency.)

4.4. Ascorbate-glutathione cycle. To allow the NADPH-monodehydroascorbate reductase reaction to function in the cycle as curated, we split the L-ascorbate peroxidase reaction (EC 1.11.1.11) into its two subreactions, which by default are not exported in the FBA problem.

4.5. Gamma-glutamyl cycle. The gamma-glutamyltransferase was lumped together with **GAMMA-GLUTAMYL-CYCLOTRANSFERASE-RXN**, originally written in terms of the instanceless class ‘L-2-AMINO-ACID’ which appeared in no other stoichiometries in the FBA export, and the dipeptidase **RXN-6622**, which is the only reaction that can consume the cysteinylglycine product of the gamma-glutamyltransferase, forming a combined reaction which can carry flux. The combined reaction retained the gene associations of the gamma-glutamyltransferase, as the other two reactions have no associated genes.

4.6. Methionine synthesis from homocysteine. The methionine synthase reaction of CornCyc’s methionine biosynthesis pathway, **HOMOCYSMET-RXN**, EC 2.1.1.14, specifically requires 5-methyltetrahydropteryltri-L-glutamate as a cofactor. Polyglutamylation of folates is present in CornCyc in an abstract representation (with tetrahydrofolate synthase catalyzing the addition of a glutamyl group to a 5-methyltetrahydropteryl with n glutamyl groups); we have not converted this into an explicit representation in the FBA model. Instead, **HOMOCYSMETB12-RXN**, EC 2.1.1.13, acts to produce methionine from homocysteine; the effects of this possible inaccuracy on the behavior of the rest of the network should be limited.

4.7. Basic import and export. The following species are given overall import/export reactions:

- WATER
- CARBON-DIOXIDE
- OXYGEN-MOLECULE
- PROTON
- NITRATE
- SULFATE
- |Pi|
- |Light|
- MG+2

These reactions exchange species inside the cell with species in meaningfully labeled compartments where possible (eg, oxygen and CO_2 are exchanged with the intercellular air space, mineral nutrients with the xylem, etc.)

In addition, to facilitate exchange among compartments in the whole-leaf model, a number of exchanges with a phloem compartment were set up: these included sucrose, glycine (as a representative of the amino acids detected in maize phloem sap by Ohshima et al [83],) and the potential phloem sulfur transport compound glutathione [84].

Note that these reactions should be inactive, or restricted to the exporting direction only, when not modeling transport within the leaf (except for sucrose, where a free supply should be allowed in heterotrophic conditions.)

4.8. Defining the biomass components. Two types of biomass reactions are added to the model:

- Sinks for individual species, for simulations (e.g, fits to RNAseq data) where the relative rates of production of different components are unknown. The species given such sinks are listed in `biomass_components.txt`.
- A set of reactions producing a combined biomass species, made up of assorted components in fixed proportions, for simulations where the maximum rate of production of biomass is of interest, and an approximately realistic biomass composition needs to be enforced directly. These reactions were taken with minor modifications from [21]; their adaptation is described below and they are listed in `adapted_irs1563.biomass.txt`.

To conceptually and practically separate these types of biomass reactions, which in general should not both be active in any one calculation, the biomass species they produce are located within two separate abstract biomass compartments in the SBML model.

In general, the biomass sink reactions have no gene associations, but an exception was made for the twenty reactions representing incorporation of amino acids into protein, which inherit the gene associations of the corresponding tRNA ligase reactions in CornCyc. (In principle these could be distinguished from sink reactions representing the expansion of free amino acid pools as cells grow and divide, but we have ignored this issue for now.)

Note that, to support the adapted iRS1563 biomass equation, a reaction representing the production of free galactose from GDP-L-galactose was introduced (otherwise, release of galactose from UDP-galactose was catalyzed by two reactions in the pathways of indole-3-acetyl-ester conjugate biosynthesis and indole-3-acetate activation, likely not a major route for carbohydrate production.) Free galactose is not included in the individual biomass species used for data fitting.

5. COMPARTMENTALIZATION

Approaches differ to the subcellular compartmentalization in FBA models of eukaryotes, ranging from the assignment of compartments to a few key pathways known to function primarily outside the cytosol, as in the mitochondrial and chloroplastic “modules” of AraMeta [109] and RiceMeta [110] to the extremely comprehensive, data-driven approach of [111]. Here, we did not attempt to comprehensively assign reactions to their proper compartments; instead, we started with a modular approach similar to [110] in which some core metabolic pathways were

compartmentalized (in our case, the TCA cycle and mitochondrial electron transport chain in the mitochondrion, the light reactions of photosynthesis, Calvin cycle, and some reactions of the C4 and photorespiratory pathways in the chlorophyll, and some reactions of the photorespiratory pathway in the peroxisome, with transport reactions added as necessary.)

We then refined the compartment assignments of other reactions and pathways as needed to permit key metabolic functions and compartmentalize a limited number of additional reactions whose incorrect assignment to the cytosol we judged particularly likely to lead to misleading results.

More details on individual compartmentalization choices and transport reactions are given below.

5.1. Intracellular transport. Sources (beyond those detailed below) informing the addition of intracellular transport reactions in the model included the transport reactions present in AraMeta [109], reviews of photorespiratory metabolism with attention to compartmentalization [112, 113], a review of chloroplast transporters [114], and a review of transport processes in C4 photosynthesis [115].

In most cases we have not tried to reflect the mechanisms of the transport systems, where those are known, in any detail (exceptions include the triose phosphate-phosphate and PEP-phosphate transporters across the chloroplast envelope), nor have we associated genes with the transporters, even when they are known. Future work should pay greater attention to this aspect of the system.

5.2. Photorespiratory pathway. Following [116] we assumed that reducing power was supplied to the peroxisome through an oxaloacetate-malate shuttle and NAD(H)-dependent malate dehydrogenase, and added an oxaloacetate-malate antiporter and a copy of MALATE-DEH-RXN to the peroxisome. Reactions of the pathway were localized following [112] and [113]. Note that glycine decarboxylase was assigned exclusively to the mitochondrion, while serine hydroxymethyltransferase was present in both the mitochondrion and the cytoplasm, where it plays a role in one-carbon metabolism [117].

5.3. Various ferredoxin-consuming pathways. The model includes several pathways or reactions (e.g., sulfite and nitrite reduction and the chlorophyll cycle) which rely on ferredoxins for reducing power, and are localized to the chloroplast, where, in the light, reduced ferredoxins may be supplied by the photosynthetic electron transport chain.

Rather than assign the reactions of these pathways to compartments appropriately, we added a reaction exchanging reduced ferredoxins and oxidized ferredoxins across the chloroplast boundary to supply ferredoxin-driven pathways in the cytosol. We emphasize that this is a convenient simplification and is not intended to represent a realistic mechanism.

5.4. Ascorbate production. The L-galactonolactone dehydrogenase responsible for the final step of the ascorbate production pathway in CornCyc reduces cytochrome C and has been experimentally localized to the mitochondrial inner membrane, with its catalytic site facing outwards, into the intermembrane space [118]. As the outer membrane is generally permeable to small molecules we have treated this reaction as acting directly on cytoplasmic galactonolactone and ascorbate. A

sink for ascorbate as a biomass component was added, as it is found in substantial quantities in leaves (see, e.g., [119, 120].)

5.5. Ascorbate-glutathione cycle. This cycle is present in multiple cellular compartments; in the model we included only cytosolic and chloroplastic instances (of which only the chloroplastic was ultimately expected to be relevant, as there was no supply of superoxides in the cytosol.) Note that none of the genes associated with monodehydroascorbate reductase could be assigned to the chloroplast under the rules described below: two had curated location in the peroxisome while GRMZM2G320307 had no curated location and TargetP prediction of mitochondrial (GRMZM2G320307_P01) and cytoplasmic (GRMZM2G320307_P02, GRMZM2G320307_P03) locations. Reduction of monodehydroascorbate may also proceed non-enzymatically (see above) so this (enzymatic) reaction was removed from the chloroplast in favor of direct reduction by ferredoxin.

6. GENE ASSOCIATIONS FOR COMPARTMENTALIZED REACTIONS

Where a reaction was present in more than one compartment— that is, when two or more reactions in different compartments were associated with the same reaction record in CornCyc— we examined the genes associated with those reactions in CornCyc and assigned them to the instance of the reaction in the most appropriate compartment, as far as possible.

Where the Plant Proteome Database [26] provided manually curated location assignments for genes, those were used; otherwise, we used automatic location predictions by TargetP [121] or in some cases referred to the gene’s annotation (both also provided by PPDB.) In general we assumed the appropriate location for a gene product was the cytoplasmic compartment absent a specific prediction of localization in the chloroplast, mitochondrion, or peroxisome. Where proteins were predicted to occur in a compartment where an no instance of a particular reaction was present, those gene associations were generally dropped from the model.

When a gene was associated with a reaction in more than one compartment and also a reaction present in only one compartment, in general the association with the reaction in only one compartment was dropped, except for reactions which we believed based on literature evidence (including comments in CornCyc and PPDB) were assigned to the cytoplasmic compartment only because our compartmentalization process was incomplete.

Some details on the judgment calls made in this process are provided in the comments to the file `gra_overrides.txt`; we comment here on a few unusual cases.

6.1. NADH dehydrogenases. Cyclic electron transport around Photosystem I may occur through the chloroplast NADH dehydrogenase complex or an alternate pathway which in Arabidopsis involves PGR5 [122, 123]. In C3 plants the PGR5-dependent pathway may play the major role in tuning the photosynthetic ATP/NADPH ratio, while the NADH dehydrogenase pathway is implicated in stress responses [123]. In contrast, in C4 plants the expression of the chloroplast NADH-dehydrogenase appears to correlate with photosynthetic ATP demand, while PGR5 expression does not, suggesting it is the NADH-dehydrogenase CET pathway which allows increased the increased ATP production required by the C4 system [124]. Thus, genes associated in CornCyc with the NADH dehydrogenase reaction for which a chloroplast location was predicted were reassociated

with the model’s cyclic electron transport reaction (despite the fact that our somewhat abstract cyclic electron transport reaction may not accurately represent the biochemistry of the NADH-dependent pathway.)

6.2. Pyruvate dehydrogenases. In practice, pyruvate dehydrogenase complexes are found in the mitochondrion and chloroplast, but here we have not fully compartmentalized the chloroplastic pyruvate dehydrogenase and the pathways it supplies, instead leaving it in the cytosol. Thus, genes associated with the reactions of the complex with predicted chloroplast localization were associated instead with the cytosolic version. Genes with no curated or predicted location were left associated with both forms (splitting their expression data between them, in the fitting process.)

7. TESTING AND CONSISTENCY CHECKING

The compartmentalized single-cell model was checked in detail for conservation violations by testing the feasibility of net production or consumption of a unit of each internal species with all external transport and biomass sink reactions suppressed.

Where such production was found feasible, the reactions involved were carefully inspected and stoichiometry coefficients adjusted to restore balance if necessary. In practice, this led only to the correction of erroneous reactions added by hand; as expected, no balance issues were found with reactions exported from CornCyc.

In the final version, no such unrealistic processes are possible in the model under normal conditions. (Note that the species representing light input may be consumed in isolation, but the use of light energy to drive a futile cycle is not unrealistic, though we have not examined the details of the process found by the consistency checker in any detail.) Of course, demonstrating that no such production/consumption is feasible does not guarantee that all reactions in the model are properly balanced.

Testing also verified that all individual biomass sink reactions, and the combined biomass reaction, could proceed at nonzero rates.

8. SBML EXPORT

8.1. Component names. SBML distinguishes a component’s name from its ID. Reactions and species in the SBML model were given name attributes according to the by calling the Pathway Tools `get_name_string` function on the frames in the database from which they derive, if any. The IDs of the SBML components were derived from the frame handles, replacing special characters with underscores as necessary to conform to the SBML `sID` standard.

Note that for some reactions in CornCyc, the result of `get_name_string` is an EC number different from the EC number indicated by the label of the frame (e.g, `2.7.1.133-RXN`, for which ‘EC 2.7.1.159’ is returned.) The frame in CornCyc (if any) from which each reaction in the SBML model is ultimately derived is preserved as a comment in the reaction’s Notes element, to resolve any ambiguity.

8.2. Gene annotations. Each reaction in the FBA model associated with a particular parent frame in CornCyc was given an association rule that combined all genes associated with that reaction in CornCyc, as well as all genes associated with

all generic reactions of which the parent reaction is a specific form, in a logical ‘or’ relationship, stored in the reaction’s Notes element per the COBRA standard.

9. MODEL REFINEMENT

9.1. Phosphoribulokinase. In early attempts to fit the model to the leaf gradient data, high costs were associated with the mesophyll phosphoribulokinase reaction in the source tissue when the bundle sheath CO_2 level was high. We noted that in CornCyc 4.0 several genes were associated with both PRK and glyceraldehyde-3-phosphate dehydrogenase. To clarify the role of these genes we referred to annotations in the Plant Proteome Database [26] and best hits in the Conserved Domain Database ([125], accessed through NCBI.) Of the eight genes associated with PRK in CornCyc, three (GRMZM2G039723, GRMZM2G337113, GRMZM2G162845) appeared to encode GAPDH enzymes (per PPDB annotations and the presence of *Gp_dh_N* and *Gp_dh_C* domains), three (GRMZM2G162529, GRMZM2G463280, GRMZM2G026024) appeared to encode genuine phosphoribulokinases (per PPDB annotations and the presence of PRK domains), and two appeared to encode CP12-type regulatory proteins, with no obvious evidence for any individual protein sharing more than one of these roles. The regulatory role of CP12 does involve forming a complex with PRK and GAPDH, but this reduces, rather than enhancing or enabling, their individual activities [126]. We removed the PRK associations of the GAPDH and CP12 genes from our model. PPDB assigned these three GAPDH genes to a plastidic location based on experimental evidence, so we associated them with those reactions exclusively (removing associations with the cytosolic instances of EC 1.2.1.13 and/or EC 1.2.1.12.)

10. BIOMASS EQUATION

We developed a biomass equation following that used in [21]. Our calculations are based on supplementary file S4 of that paper², in particular sheet 2, ‘Biomass.rxn’.

That sheet derives a biomass equation corresponding to the production of one gram of plant dry weight, based on literature data on biomass composition; the description is divided into subreactions forming (e.g.) ‘nitrogenous compounds’, ‘lignin’, etc., which then participate in an overall biomass reaction.) The units of the stoichiometric coefficients are mmol.

We have adopted most of the biomass composition assumptions of Saha et al wholesale, with gratitude for their efforts in compiling this data from the literature. However, we have made some minor adjustments, resulting in a different overall stoichiometry for biomass production.

10.1. Fatty acids. Saha et al represent the total lipid/fatty acid contribution to biomass as a pool of triglycerides in proportions apparently based on a maize oil measurement and thus probably reflective of seed triglyceride composition.

We substitute measurements of the fatty acid content of mature maize leaf membrane lipids [127] and write a biomass sub-reaction which consumes the relevant free fatty acids (rather than their derivatives in the form of triacylglycerols, membrane lipids, etc.) as shown in Table 1.

Weighting the molecular weights by the mole fractions, we find one mole of fatty

²Specifically, `journal.pone.0021784.s004.xls`, as downloaded from the PLoS One web site 20 November 2013

Fatty acid	CornCyc compound	mol. wt. (g/mol)	mole fraction
palmitic	PALMITATE	255.42	0.104
palmitoleic	CPD-9245	253.4	0.056
stearic	STEARIC_ACID	283.47	0.011
oleic	OLEATE_CPD	281.46	0.044
linoleic	LINOLEIC_ACID	279.44	0.132
linolenic	LINOLENIC_ACID	277.43	0.646

TABLE 1. Fatty acid proportions in biomass.

acid in appropriate proportions weighs 272.4 g. Dividing the mole fractions by the overall molar weight and multiplying coefficients by 1000 to convert to millimoles, we arrive at the final equation:

$$0.382 \text{ PALMITATE} + 0.206 \text{ CPD-9245} + 0.04 \text{ STEARIC_ACID} + 0.162 \text{ OLEATE_CPD} + 0.485 \text{ LINOLEIC_ACID} + 2.372 \text{ LINOLENIC_ACID} = \text{fatty_acids_biomass}$$

where the left-hand side represents 1 g.

Fractions add to less than 1.0 because we ignore trace (mole fraction ≤ 0.01) amounts of C14:0 and C20:0 fatty acids. Note that the leaf fatty acid composition is known to change along the developmental gradient, so specifying any single composition is an approximation; see [128].

10.2. Hemicellulose. We adopted the hemicellulose production reaction as is, using the species added to the model for this purpose, ‘polysaccharide_[sugar]_unit’. The resulting equation is:

$$\begin{aligned} &0.548 \text{ polysaccharide_arabinose_unit} + \\ &1.248 \text{ polysaccharide_xylose_unit} + \\ &0.301 \text{ polysaccharide_mannose_unit} + \\ &0.144 \text{ polysaccharide_galactose_unit} + \\ &3.254 \text{ polysaccharide_glucose_unit} + \\ &0.166 \text{ polysaccharide_galacturonate_unit} + \\ &0.166 \text{ polysaccharide_glucuronate_unit} = \text{hemicellulose_biomass}. \end{aligned}$$

10.3. Total carbohydrates. We recalculated the stoichiometries of the carbohydrate-producing reaction to account for the differing molecular weight of our representation of cellulose (‘CELLULOSE_monomer_equivalent’, effectively a glucose molecule), account for the fact that one unit of hemicellulose represents one gram, not one (milli)mole, and express pectin in terms of polysaccharide_galacturonate_unit, reflecting a belief that UDP is released in the formation of pectin from UDP-D-galacturonate, rather than retained in the polymer [129].

It is not clear what form the ‘mannose’ referred to by Penning de Vries et al should be assumed to take, as free mannose is not found in plants under most circumstances (see, e.g., [130–132].) Here we somewhat arbitrarily choose mannose-6-phosphate.

Table 2 shows the calculation, resulting in the equation:

$$0.067 \text{ RIBOSE} + 0.278 \text{ GLC} + 0.111 \text{ FRU} + 0.039 \text{ MANNOSE-6P} + 0.056 \text{ GALACTOSE} + 0.146 \text{ SUCROSE} + 2.220 \text{ CELLULOSE_monomer_equivalent}$$

Component	Species in model	unit wt (mg)	wt fraction	units/g product
Ribose	RIBOSE	150.053	0.010	0.067
Glucose	GLC	180.063	0.050	0.278
Fructose	FRU	180.063	0.020	0.111
Mannose	MANNOSE-6P	258.120	0.010	0.039
Galactose	GALACTOSE	180.063	0.010	0.056
Sucrose	SUCROS	342.116	0.050	0.146
Cellulose	CELLULOSE_monomer_equivalent	180.160	0.400	2.220
Hemicellulose	hemicellulose_biomass	1000.000	0.400	0.400
Pectin	polysaccharide_galacturonate_unit	193.130	0.050	0.259

TABLE 2. Carbohydrate species in biomass.

$$+ 0.400 \text{ hemicellulose_biomass} + 0.259 \text{ polysaccharide_galacturonate_unit} \\ = \text{carbohydrates_biomass}.$$

10.4. **Organic acids.** We adopt this reaction as is. In the terminology of our model, the resulting equation is:

$$0.556 \text{ OXALATE} + 0.676 \text{ GLYOX} + 1.515 \text{ OXALACETIC_ACID} + 0.746 \\ \text{MAL} + 1.562 \text{ CIT} + 1.724 \text{ CIS-ACONITATE} = \text{organic_acids_biomass}.$$

10.5. **Protein and free amino acids.** We adopt these reactions as is. In the terminology of our model, the resulting equations are:

$$1.15 \text{ L-ALPHA-ALANINE} + 0.0959 \text{ ARG} + 0.414 \text{ L-ASPARTATE} + 0.0313 \\ \text{CYS} + 1.53 \text{ GLT} + 0.0445 \text{ GLY} + 0.0915 \text{ HIS} + 0.465 \text{ ILE} + 1.51 \\ \text{LEU} + 5.71\text{e-}05 \text{ LYS} + 0.123 \text{ MET} + 0.314 \text{ PHE} + 0.762 \text{ PRO} + \\ 0.612 \text{ SER} + 0.175 \text{ THR} + 0.00409 \text{ TRP} + 0.244 \text{ TYR} + 0.25 \text{ VAL} \\ = \text{protein_biomass}$$

and

$$0.624 \text{ L-ALPHA-ALANINE} + 0.319 \text{ ARG} + 0.418 \text{ L-ASPARTATE} + 0.231 \\ \text{CYS} + 0.378 \text{ GLT} + 0.740 \text{ GLY} + 0.358 \text{ HIS} + 0.424 \text{ ILE} + 0.424 \text{ LEU} \\ + 0.380 \text{ LYS} + 0.373 \text{ MET} + 0.337 \text{ PHE} + 0.483 \text{ PRO} + 0.529 \text{ SER} + \\ 0.467 \text{ THR} + 0.272 \text{ TRP} + 0.307 \text{ TYR} + 0.475 \text{ VAL} = \text{free_aa_biomass}.$$

10.6. **Lignin.** We adopt this reaction as is. In the terminology of our model, the resulting equation is:

$$2.221 \text{ COUMARYL-ALCOHOL} + 1.851 \text{ CONIFERYL-ALCOHOL} + 1.587 \text{ SINAPYL-ALCOHOL} \\ = \text{lignin_biomass}.$$

10.7. **Nucleic acids.** We adopt this reaction as is (though note that, as discussed above, nucleotide triphosphates are not necessarily the appropriate best representation for polymerized nucleic acids). In the terminology of our model, the resulting equation is:

$$0.247 \text{ ATP} + 0.239 \text{ GTP} + 0.259 \text{ CTP} + 0.258 \text{ UTP} + 0.255 \text{ DATP} + \\ 0.247 \text{ DGTP} + 0.268 \text{ DCTP} + 0.259 \text{ TTP} = \text{nucleic_acids_biomass}.$$

10.8. **Nitrogenous compounds.** We use the same nitrogenous compound weight fraction breakdown, but recalculate the stoichiometric coefficients accounting for the fact that the protein biomass, free amino acid biomass, and nucleotide biomass species each represent one gram, so that the appropriate stoichiometric coefficients of those species for the production of one total gram of nitrogenous compounds are simply the weight fractions; see Table 3.

The resulting equation is

Component	Species in model	unit wt (mg)	wt fraction	units/g product
Amino acids	<code>free_aa.biomass</code>	1000.000	0.100	0.100
Proteins	<code>protein.biomass</code>	1000.000	0.870	0.870
Nucleic acids	<code>nucleic_acids.biomass</code>	1000.000	0.030	0.030

TABLE 3. Nitrogenous biomass breakdown.

$$0.100 \text{ free_aa.biomass} + 0.870 \text{ protein.biomass} + 0.030 \text{ nucleic_acids.biomass} \\ = \text{nitrogenous.biomass}.$$

10.9. **Inorganic materials.** We ignore these entirely, as they play no other role in the model. (Note that even in iRS1563 the two species involved, potassium and chloride, participate only in source and sink reactions.)

10.10. **Total biomass reaction.** We drop the inorganic materials term (note that weight fractions now add to 0.95) and recalculate the stoichiometric coefficients, accounting for the fact that the component biomass subspecies each represent one gram; see Table 4.

Component	Species in model	unit wt (mg)	wt fraction	units/g product
Nitrogenous compounds	<code>nitrogenous.biomass</code>	1000.000	0.230	0.230
Carbohydrates	<code>carbohydrates.biomass</code>	1000.000	0.565	0.565
Lipids	<code>fatty_acids.biomass</code>	1000.000	0.025	0.025
Lignin	<code>lignin.biomass</code>	1000.000	0.080	0.080
Organic acids	<code>organic_acids.biomass</code>	1000.000	0.050	0.050

TABLE 4. Breakdown of total biomass.

The final equation is

$$0.230 \text{ nitrogenous.biomass} + 0.565 \text{ carbohydrates.biomass} + \\ 0.025 \text{ fatty_acids.biomass} + 0.080 \text{ lignin.biomass} + 0.050 \text{ organic_acids.biomass} \\ = \text{total.biomass}.$$

Saha et al additionally incorporate an ATP cost in their overall biomass reaction, based on that used in an earlier Arabidopsis model (AraGEM [41]) Combining this ATP hydrolysis with a sink of total biomass, we arrive at the overall equation for biomass production and growth (`CombinedBiomassReaction`):

$$1.0 \text{ total.biomass} + 30.0 \text{ ATP} + 30.0 \text{ WATER} = 30.0 \text{ ADP} + 30.0 \text{ Pi} \\ + 30.0 \text{ PROTON}$$

10.11. **Protonation.** Throughout, note that the molecular weights of species in our model may differ somewhat from those used in the iRS1563 table because of differing assumptions about protonation. The practical consequences of this difference should be limited.

10.12. **Oxalate.** Early drafts of the model could not produce oxalate. CornCyc indicates its production as resulting only from ascorbic acid catabolism with concomitant production of L-threonate. Recent reviews suggest this is the primary pathway of oxalate production in plant species which form calcium oxalate crystals, with the threonate ultimately being oxidized to tartrate [133–135], though the pathways of production of soluble oxalate are less clear [134]. We found little immediate evidence that tartrate (or threonate) is formed in maize leaves at levels

comparable to that of oxalate, or of pathways which could further metabolize the tartrate.

Of the three reactions in iRS1563 which could produce oxalate, only one has an associated gene: oxalate carboxylase (oxalate = formate + CO₂); KEGG R00522 (EC4.1.1.2). The gene, 'ACG37538', may correspond to GRMZM2G103512, whose best Arabidopsis hit is AT1G09560.1 (germin-like protein 5); it may thus be more likely to be an oxalate-consuming oxidase [136] than an oxalate carboxylase, though no function was computationally predicted for GRMZM2G103512 in CornCyc.

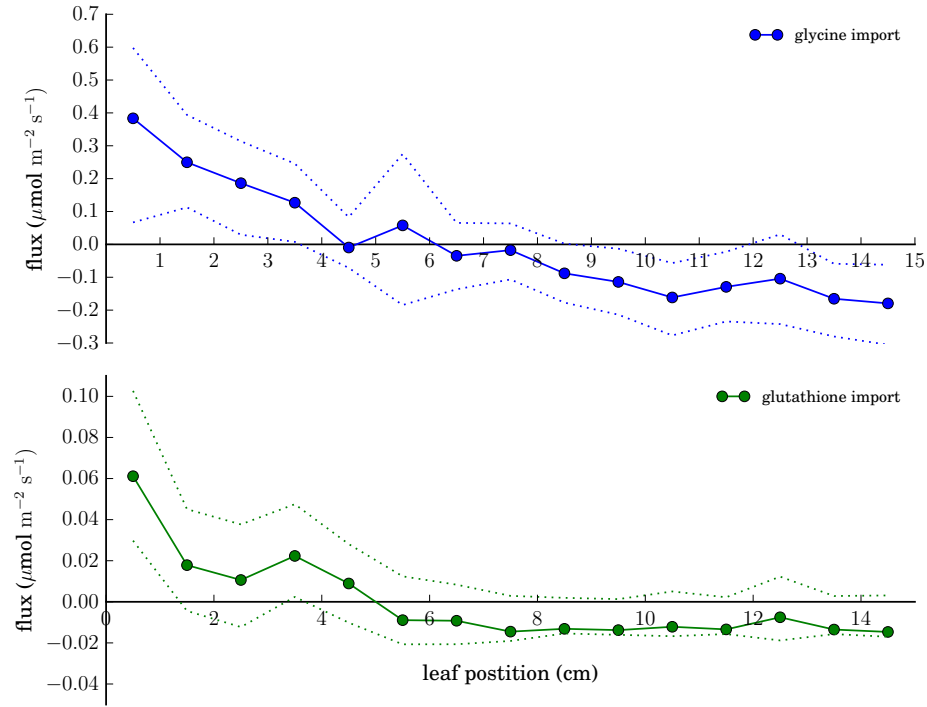
We decided the available information did not allow us to accurately model oxalate production in maize. However, to retain the iRS1563 biomass equation and ensure that mass and elemental balance was preserved, we allowed production of oxalate from oxaloacetate by oxaloacetase (EC 3.7.1.1; PlantCyc OXALOACETASE-RXN, [137]). This simple reaction has been observed in fungi [138] but is considered unlikely to be widespread in plants [134].

11. PLASMODESMATAL TRANSPORT REACTIONS

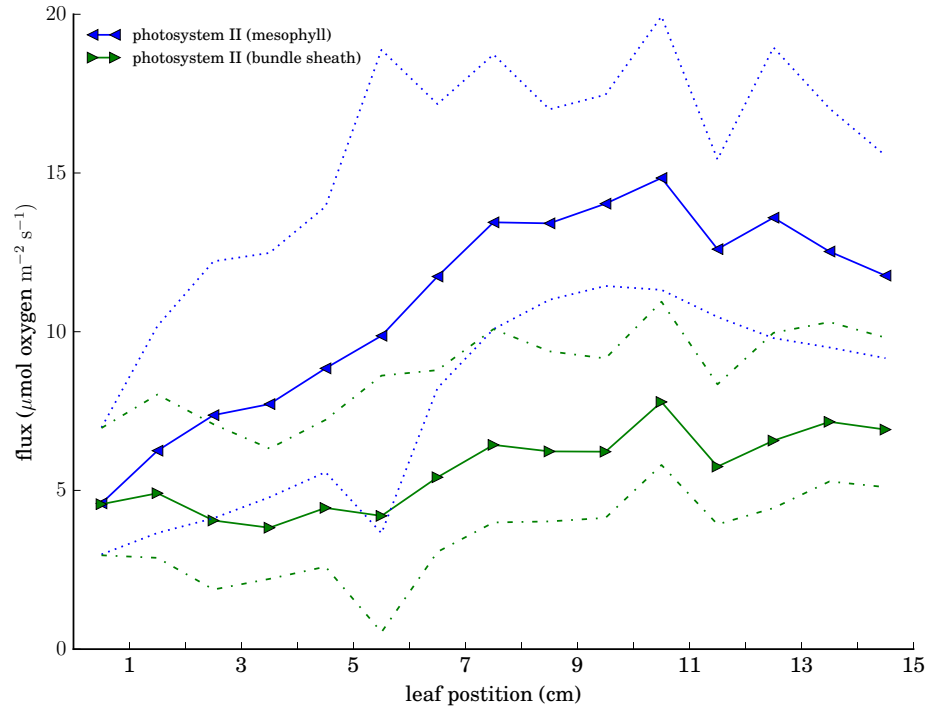
Species allowed to be exchanged between cell types through the plasmodesmata included:

- carbon dioxide and oxygen;
- known C4 cycle metabolites alanine, aspartate, malate, PEP, and pyruvate;
- the Calvin cycle intermediates glyceraldehyde 3-phosphate and 3-phosphoglycerate;
- photorespiratory metabolites glycerate, glycolate, serine, and glycine;
- nutrients sucrose, phosphate, nitrate, ammonia, sulfate and magnesium;
- glutamate and 2-ketoglutarate;
- and cysteine and glutathione [139].

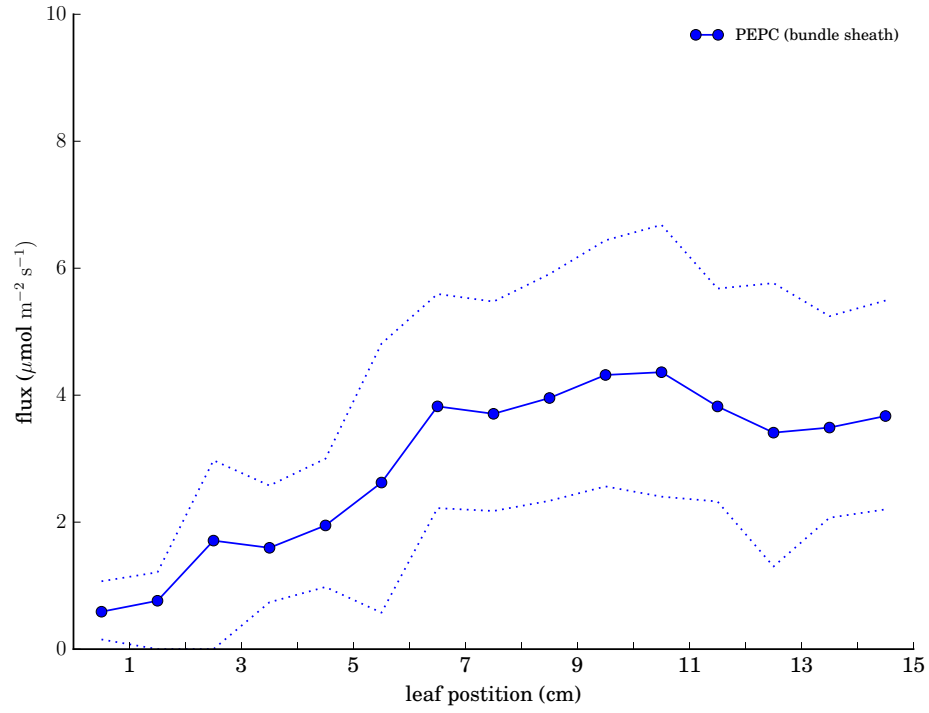
The inclusion of compounds involved in NAD-ME C4 or C3-C4 intermediate photorespiratory carbon concentrating mechanism is not meant to suggest such a system is necessarily active in maize but merely reflects our knowledge that significant transport of those species between mesophyll and bundle sheath can occur under at least some circumstances.



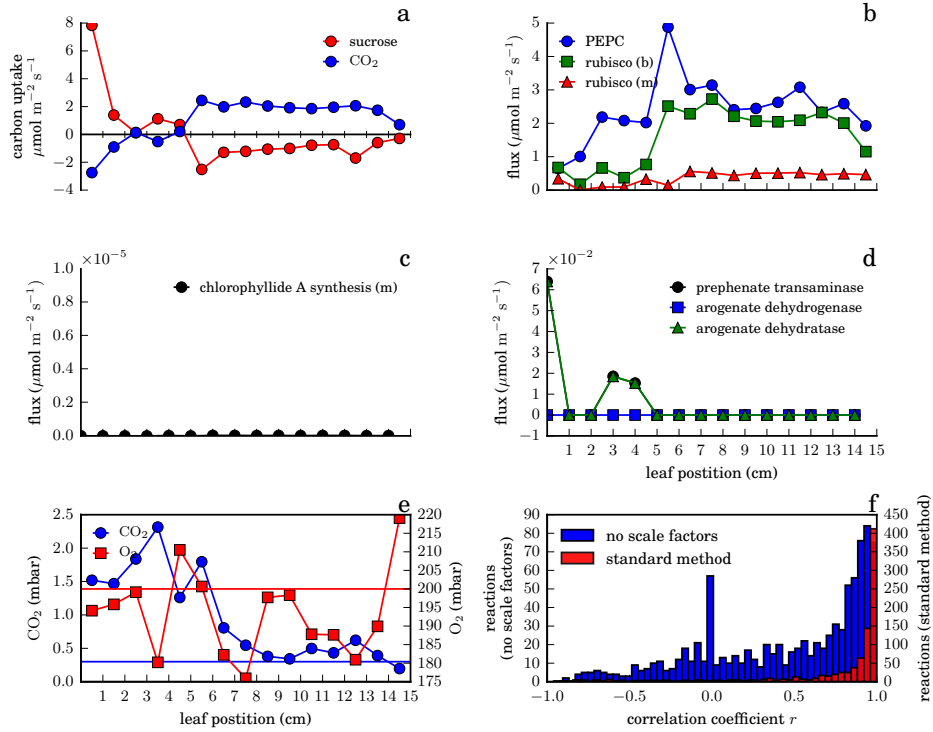
SUPPORTING FIGURE 1. **Phloem transport.** Transport of nitrogen (upper panel) and sulfur (lower panel) through the phloem in the best-fitting solution. Dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.



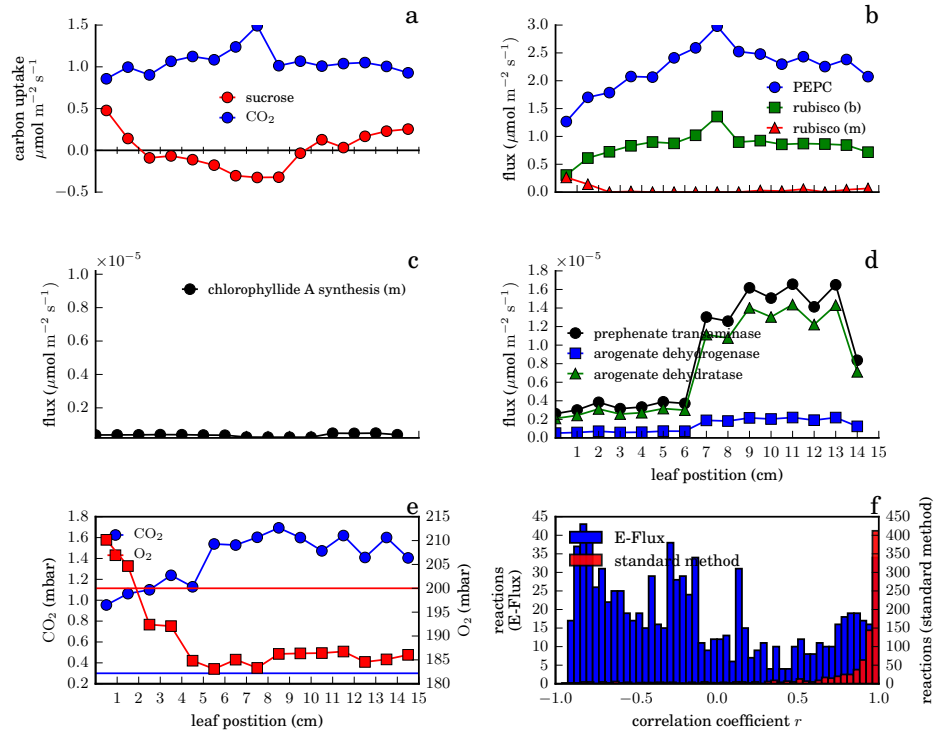
SUPPORTING FIGURE 2. **Photosystem II in mesophyll and bundle sheath.** Dashed and dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.



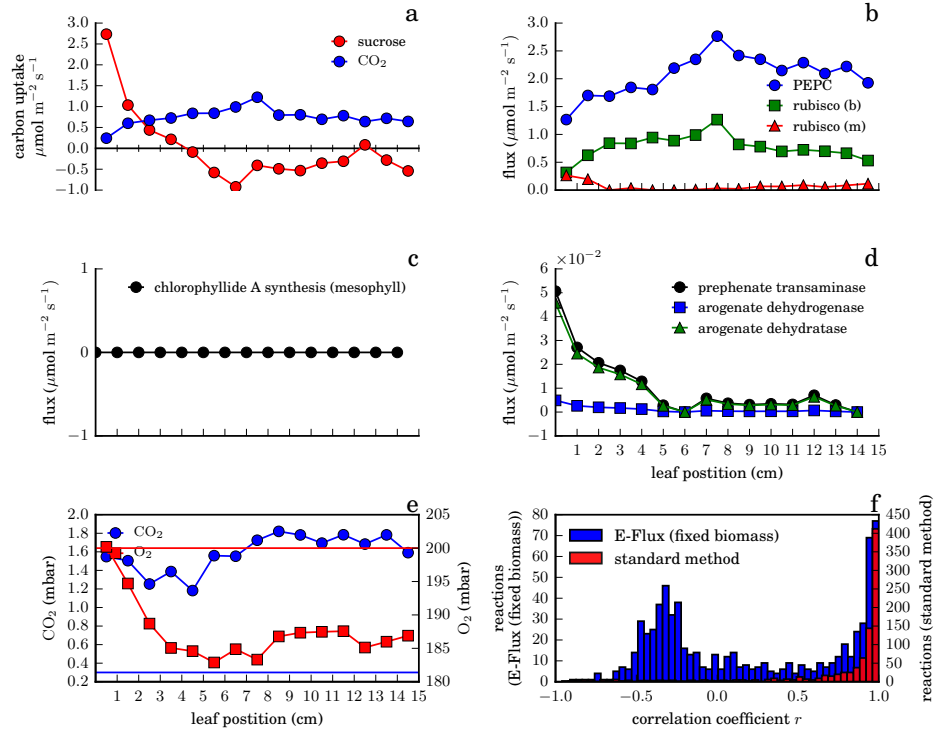
SUPPORTING FIGURE 3. **Bundle sheath PEPC flux in the best-fitting solution.** Dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.



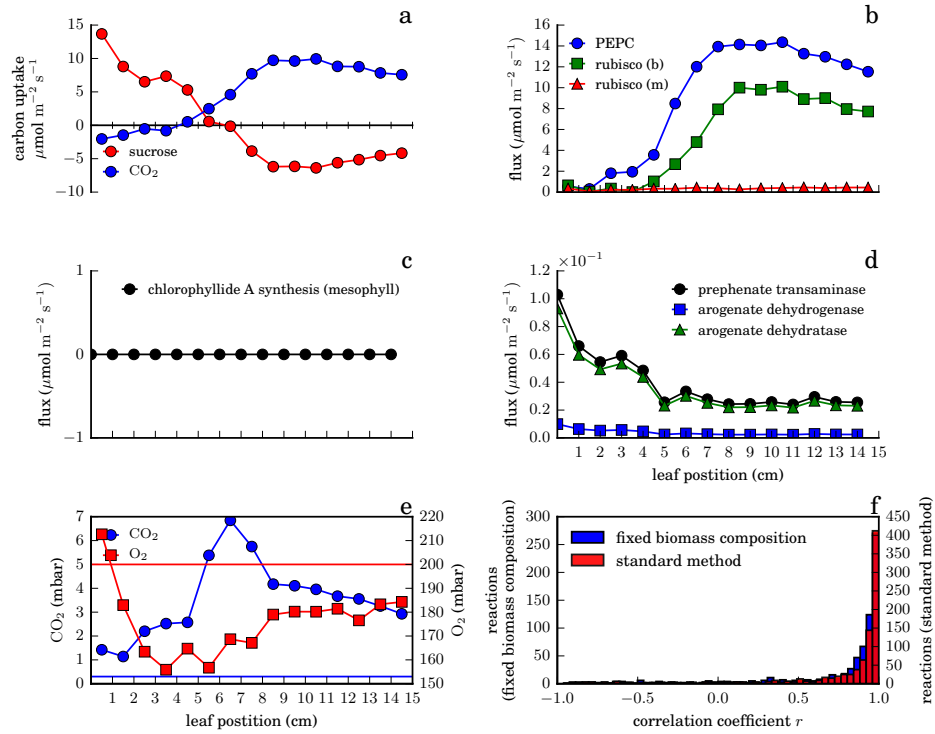
SUPPORTING FIGURE 4. Summary of predictions for the gradient model using the least-squares method without per-reaction scale factors. In eq. (8), $s_i = 0$ for all reactions i . (a) Sucrose and CO_2 uptake rates (compare to figure 3a). (b) Rates of carboxylation by PEPC and Rubisco (compare to figure 4b). (c) Predicted rate for the reactions of the chlorophyllide A synthesis pathway (compare to figure 6b.) (d) Predicted rates at the arogenate branch point (compare to figure 6d). (e) Predicted oxygen and carbon dioxide levels in the bundle sheath, with straight lines showing mesophyll levels (compare to figure 4d). (f) Distribution of correlation coefficients between data and predicted fluxes for each reaction. (blue, this method; red, standard method.) Correlation coefficients for reactions with zero predicted flux are taken to be zero, resulting in the visible peak in the histogram.



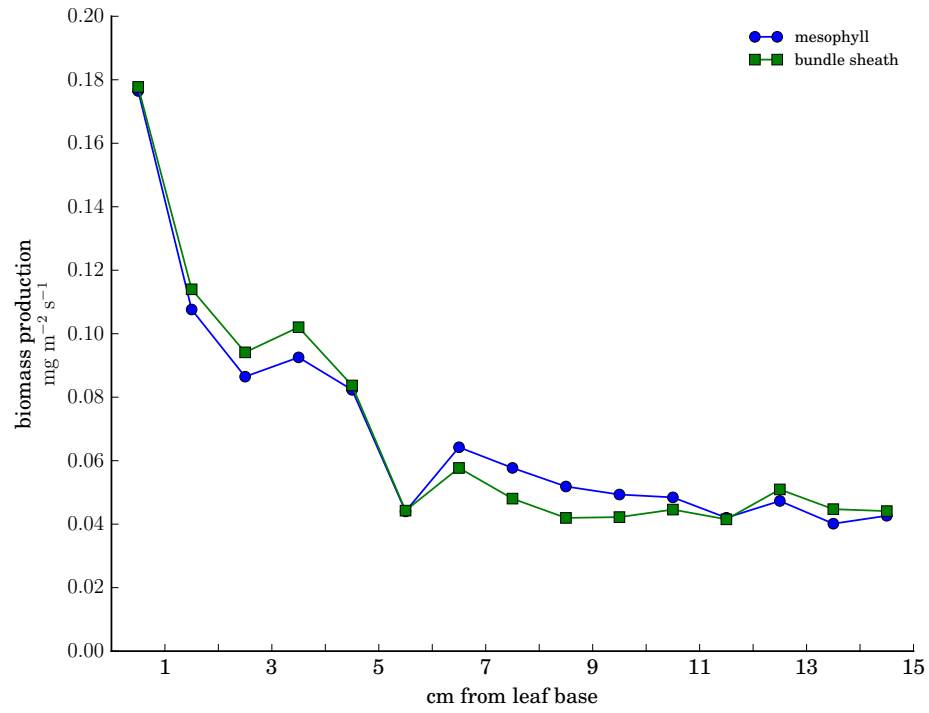
SUPPORTING FIGURE 5. Summary of predictions for the gradient model using the E-Flux method. For explanation of each panel, see Supporting Figure 4.



SUPPORTING FIGURE 6. **Summary of predictions for the gradient model using the E-Flux method with fixed biomass composition.** The biomass composition is fixed to that used by iRS1563, as adapted (see Outline). For explanation of each panel, see Supporting Figure 4. Note that the chlorophyllide A synthesis pathway is blocked when the fixed biomass composition is used.



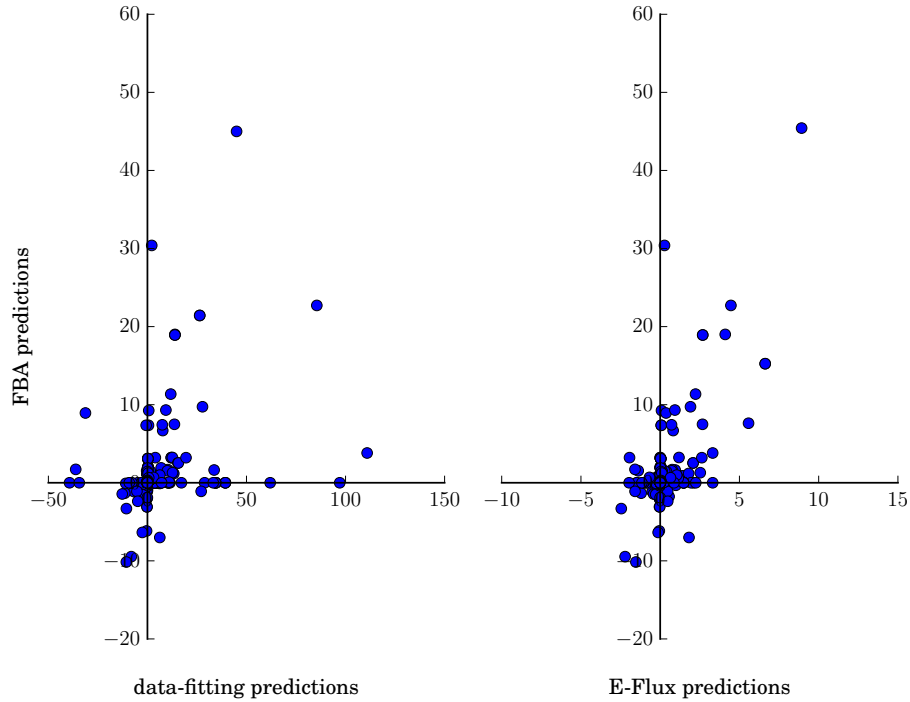
SUPPORTING FIGURE 7. **Summary of predictions for the gradient model with fixed biomass composition.** For explanation of each panel, see Supporting Figure 4. Note that the chlorophyllide A synthesis pathway is blocked when the fixed biomass composition is used.



SUPPORTING FIGURE 8. Predicted biomass production rates in mesophyll and bundle sheath cells with fixed biomass composition.

reaction	name in model	associated genes
malate dehydrogenase (NADP)	MALATE_DEHYDROGENASE_NADP_RXN_chloroplast	1
alanine aminotransferase	ALANINE_AMINOTRANSFERASE_RXN	10
aspartate aminotransferase	ASPAMINOTRANS_RXN	7
NAD-malic enzyme	EC_1.1.1.39	2
NADP-malic enzyme (cytosol)	MALIC_NADP_RXN	4
NADP-malic enzyme (chloroplast)	MALIC_NADP_RXN_chloroplast	2
PEPCK	PEPCARBOXYKIN_RXN	6
PPDK	PYRUVATEORTHOPHOSPHATE_DIKINASE_RXN_chloroplast	2
adenylate kinase	ADENYL_KIN_RXN_chloroplast	6
pyrophosphatase	INORGPYROPHOSPHAT_RXN_chloroplast	2

SUPPORTING TABLE 1. Detailed parameters contributing to the effective PEP regeneration rate: reactions in the genome-scale model which contribute to the effective maximum PEP regeneration capacity, and the number of genes associated with each. In addition to the reactions listed, transport capacities of pyruvate, PEP, alanine, aspartate and malate across the plasmodesmata and pyruvate, PEP, malate and oxaloacetate across the chloroplast inner membrane could limit this rate; the model currently associates no genes with these transport reactions.



SUPPORTING FIGURE 9. **Predicted variable values in an FBA calculation that does not incorporate expression data, compared to the best-fit and E-Flux methods.** The FBA calculation minimizes total flux while achieving the same total rate of CO_2 assimilation as predicted at the tip of the leaf in the fitting results. Left panel, FBA reaction rates vs. reaction rates predicted at the tip of the leaf in the best-fitting solution; right panel, FBA reaction rates vs. reaction rates predicted at the tip of the leaf by the E-Flux method. Axis limits exclude a small number of reactions of particularly large flux. Fluxes in $\mu\text{mol m}^{-2} \text{s}^{-1}$.

INDEX TO ADDITIONAL SUPPORTING INFORMATION FILES

Except as noted, these are available as arXiv ancillary files.

S11 Model. iEB5204 in SBML format.

S12 Model. iEB2140 in SBML format.

S13 Model. iEB2140x2 in SBML format.

S14 Protocol. Source code for the nonlinear constraint-based modeling package fluxtools. Available at <http://github.com/ebogart/fluxtools>.

S15 Protocol. Source code and input files for the calculations discussed above. Available at http://github.com/ebogart/multiscale_c4_source.

S16 Table. Predicted variable values along the leaf gradient.

S17 Table. Upper and lower bounds on predicted values of selected variables along the leaf gradient, from FVA calculations.